11-1-2017

Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory

Anne Chao
National Tsing Hua University

Chun Huo Chiu
National Tsing Hua University

Robert K. Colwell
University of Connecticut

Luiz Fernando S. Magnago
Universidade Federal de Lavras

Robin L. Chazdon
University of Connecticut

See next page for additional authors

Recommended Citation

Chao A, Chiu CH, Colwell RK, Magnago LF, Chazdon RL, Gotelli NJ. Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory. Ecology. 2017 Nov;98(11):2914-29.

Ecology, 98(11), 2017, pp. 2914–2929

Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory

Anne Chao,[1,8] Chun-Huo Chiu,[1] Robert K. Colwell,[2,3,4] Luiz Fernando S. Magnago,[5] Robin L. Chazdon,[2,6] and Nicholas J. Gotelli[7]

[1]*Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043 Taiwan*
[2]*Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, Connecticut 06269 USA*
[3]*University of Colorado Museum of Natural History, Boulder, Colorado 80309 USA*
[4]*Departmento de Ecologia, Universidade Federal de Goiás, CP 131, 74.001-970, Goiânia, GO, Brasil*
[5]*Departamento de Biologia, Setor de Ecologia e Conservação, Universidade Federal de Lavras, Lavras 37200-000 Brasil*
[6]*Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, Colorado 80309 USA*
[7]*Department of Biology, University of Vermont, Burlington, Vermont 05405 USA*

Abstract. Estimating the species, phylogenetic, and functional diversity of a community is challenging because rare species are often undetected, even with intensive sampling. The Good-Turing frequency formula, originally developed for cryptography, estimates in an ecological context the true frequencies of rare species in a single assemblage based on an incomplete sample of individuals. Until now, this formula has never been used to estimate undetected species, phylogenetic, and functional diversity. Here, we first generalize the Good-Turing formula to incomplete sampling of two assemblages. The original formula and its two-assemblage generalization provide a novel and unified approach to notation, terminology, and estimation of undetected biological diversity. For species richness, the Good-Turing framework offers an intuitive way to derive the non-parametric estimators of the undetected species richness in a single assemblage, and of the undetected species shared between two assemblages. For phylogenetic diversity, the unified approach leads to an estimator of the undetected Faith's phylogenetic diversity (*PD*, the total length of undetected branches of a phylogenetic tree connecting all species), as well as a new estimator of undetected *PD* shared between two phylogenetic trees. For functional diversity based on species traits, the unified approach yields a new estimator of undetected Walker et al.'s functional attribute diversity (*FAD*, the total species-pairwise functional distance) in a single assemblage, as well as a new estimator of undetected *FAD* shared between two assemblages. Although some of the resulting estimators have been previously published (but derived with traditional mathematical inequalities), all taxonomic, phylogenetic, and functional diversity estimators are now derived under the same framework. All the derived estimators are theoretically lower bounds of the corresponding undetected diversities; our approach reveals the sufficient conditions under which the estimators are nearly unbiased, thus offering new insights. Simulation results are reported to numerically verify the performance of the derived estimators. We illustrate all estimators and assess their sampling uncertainty with an empirical dataset for Brazilian rain forest trees. These estimators should be widely applicable to many current problems in ecology, such as the effects of climate change on spatial and temporal beta diversity and the contribution of trait diversity to ecosystem multi-functionality.

Key words: functional attribute diversity; functional diversity; phylogenetic diversity; shared diversity; species diversity; taxonomic diversity.

Introduction

Nearly all biodiversity studies and analyses are based on sampling data taken from focal assemblages. However, due to practical limitations, it is virtually impossible to detect all species, especially in hyperdiverse assemblages with many rare species. In almost every biodiversity survey and monitoring project, some proportion of the species that are present fail to be detected. Not only the presence, but also the functional traits of these species remain undetected. Moreover, the placement of undetected species on the phylogenetic tree of the observed species is unknown. Consequently, traditional measures of species, functional, and phylogenetic diversity from sample data typically underestimate the true diversities (observed plus undetected). The magnitude of this negative bias can be substantial.

For species diversity, the estimation of undetected richness based on incomplete samples from a single assemblage has been widely applied, not only in ecology and conservation biology, but also in many other disciplines; see Colwell and Coddington (1994), Chazdon et al. (1998), Magurran (2004), Chao (2005), Hortal

Manuscript received 21 September 2016; revised 27 April 2017; accepted 24 July 2017. Corresponding Editor: Tom E. X. Miller.
[8] E-mail: chao@stat.nthu.edu.tw

et al. (2006), Gotelli and Colwell (2011), Gotelli and Chao (2013) and Chao and Chiu (2016) for various applications. For two assemblages, shared species richness plays an important role in assessing assemblage overlap and forms a basis for constructing various types of beta diversity and (dis)similarity measures, such as the classic Sørensen and Jaccard indices (Colwell and Coddington 1994, Magurran 2004, Jost et al. 2011, Gotelli and Chao 2013). Compared with estimating species richness in a single assemblage, the estimation of shared species richness, taking undetected species into account, has received relatively little attention; see Chao and Chiu (2012) for a review.

In traditional measures of species diversity, all species (or taxa at some other rank) are considered to be equally distinct from one another. Species differences can be based directly on their evolutionary histories, either in the form of taxonomic classification or well-supported phylogenetic trees. A rapidly growing literature addresses phylogenetic diversity metrics and related (dis)similarity measures; see Cavender-Bares et al. (2012) for a review. A widely used phylogenetic metric is Faith's (1992) *PD* (phylogenetic diversity), which is defined as the sum of the branch lengths of a phylogenetic tree connecting all species in the target assemblage. Throughout this paper, *PD* refers to Faith's (1992) *PD*. For most data sets, *PD* is highly correlated with species richness (e.g., Matos et al. 2017). When a sample fails to detect all species present, the lineages/branches associated with these undetected species are also missing from the phylogenetic tree of the observed species. The undetected *PD* in an incomplete sample was not discussed until recent years (Cardoso et al. 2014, Chao et al. 2015*a*).

The phylogenetic version of the Jaccard dissimilarity index is referred to as the *UniFrac* measure, developed by Lozupone and Knight (2005). The phylogenetic version of the Sørensen similarity index is referred to as the *PhyloSør* (phylo-Sørensen) index, developed by Bryant et al. (2008) and Ferrier et al. (2007). All these phylogenetic (dis)similarity measures are based on the shared branch lengths between two phylogenetic trees. However, to our knowledge, the estimation of the undetected shared *PD* (i.e., the total length of undetected branches shared by two phylogenetic trees) has not previously been discussed in the literature.

When species are described by a set of traits that affect organismal and/or ecosystem functioning, pairwise species differences within an assemblage can also be measured by the dissimilarity or distances between their trait profiles, which can be weighted or unweighted by their abundances. Functional diversity or trait diversity quantifies the diversity of species' traits among coexisting species in an assemblage (Tilman et al. 1997, Diaz and Cabido 2001, Swenson et al. 2012). Functional diversity is regarded as key to understanding ecosystem processes and their response to environmental stress or disturbance (Cadotte et al. 2009). A distance-based measure at the assemblage level for quantifying functional diversity is *FAD* (functional attribute diversity, as defined by Walker et al. 1999), which is the sum of the species-pairwise functional distances. A modified version called *MFAD* (modified *FAD*) was proposed by Schmera et al. (2009), who replaced species with "functional units" (here the collection of all species with identical traits is regarded as a single functional unit). For simplicity, we focus on the estimation of *FAD*, but a similar approach can be applied if species are replaced by functional units or other clusters of species. For incomplete samples, the traits of undetected species are missing, and thus their pairwise distances are not recorded and cannot be considered in the observed *FAD*. As far as we are aware, there have been no estimators previously developed for undetected *FAD* in a single assemblage or undetected *FAD* shared by two assemblages.

In his famous cryptanalysis to crack German ciphers during World War II, Alan Turing, regarded as the founder of modern computer sciences, developed novel statistical methods to estimate the true frequencies of rare code elements (including still-undetected code elements), based on the observed frequencies in "samples" of intercepted Nazi code. According to Good (1953, 2000), Turing never published his wartime statistical work, but permitted Good to publish it after the war. The two influential papers by Good (1953) and Good and Toulmin (1956) presented Turing's wartime statistical work on the frequency formula and related topics. The frequency formula is now referred to as the Good-Turing frequency formula, which has a wide range of applications in biological sciences, statistics, computer sciences, information sciences, and linguistics, among others (McGrayne 2011, p. 100).

In an ecological context, Turing's statistical problem can be formulated as an estimation of the true frequencies of rare species when a random sample of individuals is drawn from an assemblage. In Turing's case, the species abundances are highly heterogeneous, with many rare species, so that all samples have undetected species. The Good-Turing formula answers the following question: given a species that appears r times ($r = 0, 1, 2, \ldots$) in an incomplete sample of n individuals, what is its true relative frequency in the entire assemblage? As will be described below, Turing gave a surprising answer that is contrary to most people's intuition.

Until now, the Good-Turing formula has never been applied to estimate undetected species, phylogenetic, and functional diversity. In this paper, we generalize the Good-Turing formula, originally developed for a single assemblage, to two assemblages. We also extend the Good-Turing formula to a phylogenetic version that incorporates evolutionary history among species as well as a functional version that takes into account functional traits associated with each species. We show in this paper that the Good-Turing formula and its generalizations can be used in a unified way to derive estimators of undetected species, phylogenetic, and functional diversities based on incomplete samples.

For species diversity, we apply the Good-Turing formula to intuitively derive an estimator of the number of undetected species in an assemblage. The resulting estimator turns out to be the Chao (1984) non-parametric lower bound. The two-assemblage generalized formula yields Pan et al.'s (2009) lower bound of the number of undetected shared species when a sample of individuals is taken from each of two assemblages. For phylogenetic diversity, the unified approach yields a recently published estimator of undetected *PD* in a single assemblage (Cardoso et al. 2014, Chao et al. 2015*a*). The two-assemblage generalized formula yields a new estimator of the undetected *PD* shared between two assemblages. For functional diversity based on species traits, the formulas yield a new estimator of undetected *FAD* in a single assemblage as well as a new estimator of undetected *FAD* shared between two assemblages.

All the resulting Good-Turing estimators of undetected biodiversity measures (including three previously published, derived from traditional mathematical inequalities, and three new ones) are theoretically lower bounds of the corresponding undetected diversities. Good-Turing's perspectives also reveal the sufficient conditions under which the derived estimators are nearly unbiased. It would be impossible to reveal these conditions with the traditional derivations of the corresponding estimators from inequalities. Thus, Good-Turing perspectives further justify and provide new insights for the three established estimators.

An important advance beyond previous studies is that Good-Turing's perspective here provides a novel and unified approach to the notation, terminology, and estimation of undetected diversity. All taxonomic, phylogenetic and functional diversity estimators are now derived and linked under the same framework. Simulation results are reported here to numerically validate Good-Turing theory and examine the performance of the derived estimators. An empirical dataset for tree species, collected in Brazilian rain forest, is used to illustrate all estimators.

Species Diversity: One-Assemblage Good-Turing Formula

The original Good-Turing formula

In the one-assemblage model formulation, we assume that there are S species and their true species relative abundances (or generally, species detection probabilities) are denoted by $(p_1, p_2, \ldots, p_S)$, $\sum_{i=1}^{S} p_i = 1$. Assume a sample of n individuals is selected with replacement. Here we follow Good-Turing's original model (Good 1953) in which the detection probability of each species is simply its relative abundance; all our derivations can be directly extended to a general model in which species detection probability is proportional to the product of its abundance and individual detectability; see *Discussion*. Let X_i denote the species frequency (abundance) of the i-th species in the sample, $i = 1, 2, \ldots, S$, $\sum_{X_i>0} X_i = n$. Only species with frequency $X > 0$ in the sample are detected. Those species with frequency $X = 0$ in the sample are not detected (although they are present in the assemblage) and are therefore not included in the sample data.

We define the *abundance frequency count* f_r as the number of species each represented by exactly r individuals in the sample. Good (1953) referred to it as "the frequency of frequency r." Thus f_1 is the number of "singletons" (those species that are represented by exactly 1 individual in the sample), and f_2 is the number of "doubletons" (those that are represented by exactly 2 individuals in the sample). Let S_{obs} denote the total number of those species observed in the sample; $S_{obs} = \sum_{i>0} f_i \equiv f_+$. Also, f_0 is the number of undetected species: species that are present in the assemblage of S species, but were not detected in the sample of n individuals and S_{obs} species. Therefore, we have $S_{obs} + f_0 = S$.

Turing's statistical problem can be formulated as follows. Given data, for those species that each appeared exactly r times ($r = 0, 1, 2, \ldots$) in an incomplete sample of n individuals, the mean of their true relative abundances/frequencies in the assemblage, α_r, can be mathematically expressed as

$$\alpha_r = \sum_{i=1}^{S} p_i\, I(X_i = r)/f_r, \quad r = 0, 1, 2, \ldots \tag{1a}$$

where $I(A)$ is the indicator function, i.e., $I(A) = 1$ if the event A occurs, and 0 otherwise. The numerator in Eq. (1a) represents the total true relative frequencies of those species that each appeared exactly r times in the sample. Dividing the total by f_r, we obtain the mean (per species) of their relative frequencies. Turing and Good focused on the case of small r, i.e., rare species (or rare code elements, in Turing's case). Note that for the special case of $r = 0$, Eq. (1a) implies

$$\alpha_0 f_0 = \sum_{i=1}^{S} p_i I(X_i = 0), \tag{1b}$$

which is the "coverage deficit" (Chao and Jost 2012) or the complement of the "sample coverage" defined in Good (1953). The coverage deficit of the sample quantifies the proportion of the total individuals in the assemblage that belong to undetected species; it is also the probability that a new, previously-undetected species would be found if the sample were enlarged by one individual. This is a very important measure in diversity estimation (Chao and Jost 2012).

Turing and Good discovered a surprisingly simple and remarkably effective, although non-intuitive, estimator for α_r. The Good-Turing frequency formula states that α_r, $r = 0, 1, 2, \ldots$, is not estimated by its sample frequency r/n, but rather by

$$\tilde{\alpha}_r = \frac{(r+1)}{n}\frac{f_{r+1}}{f_r}, \quad r = 0, 1, 2, \ldots \tag{1c}$$

In other words, α_r should be estimated by r^*/n, where $r^* = (r+1)f_{r+1}/f_r$. The Good-Turing frequency formula is thus contrary to most people's intuition because the estimator in (1c) depends not only on the sample frequency r of the focal species, but also on the frequency information derived from species in the next frequency class, $r+1$.

Good (1953) used a fully Bayesian approach to theoretically justify the formula (1c), whereas Robbins (1968) derived it as an empirical Bayes estimator. Good (2000) wrote "when preparing my 1953 article, I had forgotten Turing's somewhat informal proof in 1940 or 1941, which involved cards or urn models in some way, and I worked out a separate proof [Bayes estimator]. I still don't recall Turing's proof." Nevertheless, Good (1983, p. 28) provided a very intuitive non-Bayesian justification of the Good-Turing frequency formula as follows: Given an original sample of size n, consider the probability of the event that the next individual will be a species that had appeared r times in the original sample. (Mathematically, this probability is simply $\sum_{i=1}^{S} p_i I(X_i = r) = \alpha_r f_r$, as defined in Eq. 1a.) If this event occurs, then the species to which the additional individual belongs must appear $r+1$ times in the enlarged sample of size $n+1$. Because the order in which individuals were sampled is assumed to be irrelevant, the total number of individuals in the enlarged sample of size $n+1$ for those species (that appeared in the additional individual and had appeared r times in the original sample) is $(r+1)f_{r+1}$. Thus, the probability of the aforementioned event in the enlarged sample of size $n+1$ is $(r+1)f_{r+1}/(n+1)$, which can be approximated by $(r+1)f_{r+1}/n$ if n is large enough. Dividing this by the number of such species, f_r, we obtain the mean relative abundance of those species, which is the classic Good-Turing frequency formula as given in Eq. (1c). Chiu et al. (2014*b*) proposed an improved formula $\hat{\alpha}_r$ shown below for $r = 0, 1, 2, \ldots$,

$$\hat{\alpha}_r = \frac{(r+1)f_{r+1}}{(n-r)f_r + (r+1)f_{r+1}} \approx \frac{(r+1)f_{r+1}}{(n-r)f_r}. \qquad (1d)$$

This improved estimator generally has smaller mean squared error than the original Good-Turing estimator. In our subsequent derivation, we adopt the rightmost term in Eq. (1d); a simple non-Bayesian proof is provided (in Appendix S1) to facilitate the generalization to the two-assemblage case.

Undetected species richness

Statistically, species richness (observed species plus the number of undetected species) is difficult to estimate accurately if there are many almost undetectable species in a hyper-diverse community. Practically, an accurate lower bound for species richness is preferable to an inaccurate point estimator. We now demonstrate that the improved Good-Turing formula can be intuitively used to provide a lower bound for the number of undetected species and to clearly reveal, for the first time, the conditions under which the lower bound is a nearly unbiased point estimator.

For $r = 0$, both Eq. (1c) and Eq. (1d) imply that the mean population relative frequency for those undetected species is approximately $\hat{\alpha}_0 = f_1/(nf_0)$, which is not obtainable from observed data because f_0 is unknown. However, this relation implies that the product of α_0 and f_0, the estimated proportion of the total number of individuals that is due to undetected species, can be well estimated by the proportion of singletons, f_1/n. For notational simplicity, let $\widehat{\alpha_0 f_0}$ denote the estimator of the product of α_0 and f_0. Then we have

$$\widehat{\alpha_0 f_0} = \frac{f_1}{n}. \qquad (2a)$$

Equation (1d) also implies that, for those species that appeared as singletons ($r = 1$) in a sample, their mean relative frequency is estimated by

$$\hat{\alpha}_1 = \frac{2f_2}{(n-1)f_1}. \qquad (2b)$$

Intuitively, we expect that the mean relative frequency of all undetected species should be less than the mean relative frequency of all singletons, i.e., $\alpha_0 \le \alpha_1$, and this ordering should be preserved by the corresponding estimates. Combining (2a) and (2b), we readily obtain a lower bound for the number of undetected species:

$$\hat{f}_0 = \frac{\widehat{\alpha_0 f_0}}{\hat{\alpha}_0} \ge \frac{\widehat{\alpha_0 f_0}}{\hat{\alpha}_1} = \frac{\frac{f_1}{n}}{\frac{2f_2}{(n-1)f_1}} = \frac{(n-1)}{n}\frac{f_1^2}{2f_2}. \qquad (2c)$$

This lower bound for f_0 is identical to that proved rigorously by Chao (1984, 1987) by means of a Cauchy-Schwarz inequality: $E(f_0) \times 2E(f_2) \ge (1 - 1/n)[E(f_1)]^2$, which may not be intuitively understood by most ecologists. Here Good-Turing's approach is intuitive and provides a sufficient condition for the resulting estimator being unbiased, as elaborated later. Based on Eq. (2c), the estimated number of undetected species is based exclusively on the information on the rarest observed species (the number of singletons and doubletons). The idea behind this lower bound is that detected abundant species carry negligible information about the undetected species; detected rare species carry nearly all such information. From Eq. (2c), the Good-Turing formula leads to the following Chao1 species richness estimator, with a slight modification when $f_2 = 0$. (Colwell and Coddington 1994 gave the name *Chao1* to this estimator):

$$\hat{S}_{Chao1} = \begin{cases} S_{obs} + \dfrac{(n-1)}{n}\dfrac{f_1^2}{2f_2}, & \text{if } f_2 > 0, \\ S_{obs} + \dfrac{(n-1)}{n}\dfrac{f_1(f_1-1)}{2}, & \text{if } f_2 = 0. \end{cases} \qquad (3a)$$

Notice that, in the above derivation, if $\hat{\alpha}_0 \approx \hat{\alpha}_1$ (i.e., undetected species and singletons have identical mean relative abundances), then the inequality sign in Eq. (2c) becomes an equality sign, implying that the lower bound becomes an unbiased point estimator. Only through the Good-Turing perspectives can this condition be revealed.

A simple sufficient condition for the Chao1 lower bound being nearly unbiased is that *rare* species (specifically, singletons and undetected species) have approximately homogenous abundances (because this implies $\hat{\alpha}_0 \approx \hat{\alpha}_1$); in this case, the abundant species could be highly heterogeneous without affecting the estimator. However, if *rare* species are heterogeneous and sample size is not sufficiently large, then at best we can provide only a lower bound, because the data provide insufficient information to accurately estimate diversity when rare species abundances are substantially heterogeneous. A simulation study is described to examine the performance of the Chao1 estimator in *Discussion*.

Chao (1987) applied a standard approximation method and obtained the following estimated variance estimators:

$$\hat{var}(\hat{S}_{Chao1}) = f_2 \left[\frac{1}{4}\left(\frac{n-1}{n}\right)^2 \left(\frac{f_1}{f_2}\right)^4 + \left(\frac{n-1}{n}\right)^2 \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{2}\left(\frac{n-1}{n}\right)\left(\frac{f_1}{f_2}\right)^2 \right] \quad (3b)$$

A confidence interval for species richness can then be obtained using a log-transformation, so that the lower confidence bound is always greater than the observed species richness if sampling is incomplete (i.e., there are singletons in the sample). When $f_2 = 0$, a slight modification of the variance formula is needed; see Chao and Chiu (2012).

Species Diversity: Two-Assemblage Good-Turing Formula

Two-assemblage Good-Turing formulas

We now extend the one-assemblage model formulation and data framework to two assemblages (I and II), which can differ not only in their species richness, but also in their species composition. Assume that there are S species in the *pooled* assemblage. The true relative species abundances or frequencies in Assemblages I and II are denoted by $(p_{11}, p_{21}, \ldots, p_{S1})$ and, $(p_{12}, p_{22}, \ldots, p_{S2})$ respectively, $p_{i1}, p_{i2} \geq 0, \ i = 1, 2, \ldots, S$. Let the number of shared species between the two assemblages be S_{shared} (or S_{12} for notational simplicity). Without loss of generality, we assume that the first S_{shared} species in the pooled assemblage are the shared species.

A random sample is taken from each of the two assemblages (Sample I with size n_1 from Assemblage I and Sample II with size n_2 from Assemblage II). Denote the number of observed shared species by $S_{shared,obs}$, and the observed species abundances in the two assemblages, respectively, by $(X_{11}, X_{21}, \ldots, X_{S1})$ and $(X_{12}, X_{22}, \ldots, X_{S2})$. For any two non-negative integers r and v, define

$$f_{rv} = \sum_{i=1}^{S_{12}} I(X_{i1} = r, \ X_{i2} = v), \ r, v = 0, 1, 2, \ldots \quad (4a)$$

That is, f_{rv} denotes the number of *shared* species that are observed r times in Sample I and v times in Sample II. In particular, f_{11} denotes the number of shared species that are singletons in both samples, and f_{00} denotes the number of shared species that are undetected in both samples. Also, let f_{r+} denote the number of shared species that are observed r times in Sample I and are observed at least once (using a "+" sign to replace the index v) in Sample II, with a similar symmetric definition for f_{+v}. Thus, f_{++} becomes the number of observed species shared between the two samples. Mathematically, we have the following expressions:

$$f_{r+} = \sum_{i=1}^{S_{12}} I(X_{i1} = r, \ X_{i2} > 0) = \sum_{v>0} f_{rv}, \quad (4b)$$

$$f_{+v} = \sum_{i=1}^{S_{12}} I(X_{i1} > 0, \ X_{i2} = v) = \sum_{r>0} f_{rv}, \quad (4c)$$

$$f_{++} = \sum_{i=1}^{S_{12}} I(X_{i1} > 0, \ X_{i2} > 0) = \sum_{r,v>0} f_{rv} = S_{shared,obs}. \quad (4d)$$

Here we generalize the original Good-Turing formula to two assemblages. There are two parts to the generalization. The first part, below, is a direct generalization of the original formula; the second part is proved in Appendix S1 by an argument parallel to that applied in the developing the original formula.

(1) Given two-sample data, let $\alpha_{r+} = \sum_{i=1}^{S_{12}} p_{i1} I(X_{i1} = r, \ X_{i2} > 0)/f_{r+}$ be the mean of the true relative frequencies in Assemblage I for those shared species that each appeared exactly r times in Sample I and appeared at least once in Sample II. A direct generalization of the original Good-Turing formula in Eq. (1d) leads to

$$\hat{\alpha}_{r+} = \frac{(r+1)f_{r+1,+}}{(n_1 - r)f_{r+}}, \ r = 0, 1, 2, \ldots \quad (5a)$$

Similarly, we have a symmetric formula for the mean of the true relative frequencies in Assemblage II for those shared species that appeared at least once in Sample I and appeared v times in Sample II.

$$\hat{\alpha}_{+v} = \frac{(v+1)f_{+,v+1}}{(n_2 - v)f_{+v}}, \ v = 0, 1, 2, \ldots \quad (5b)$$

(2) For any shared species that appeared exactly r times in each sample, consider calculating the product of its true relative abundances in the two assemblages; the

mean of the products among all such shared species (there are f_{rr} such shared species) can be expressed as $\alpha_{rr} = \sum_{i=1}^{S_{12}} p_{i1}\, p_{i2}\, I(X_{i1} = r,\ X_{i2} = r)/f_{rr},\ \ r = 0,\ 1,\ 2\ldots$ The following generalized two-assemblage Good-Turing formula provides an estimator for α_{rr} (see Appendix S1 for a proof):

$$\hat{\alpha}_{rr} = \frac{(r+1)^2 f_{r+1,r+1}}{(n_1 - r)(n_2 - r) f_{rr}}, r = 0, 1, 2, \ldots \tag{5c}$$

The generalized formulas in Eqs. (5a)–(5c) lead elegantly to an estimator of undetected shared species richness between two assemblages, as shown below.

Undetected shared species richness between two assemblages

Under the two-assemblage model formulation and data framework, the true number of shared species can be expressed as the sum of four terms:

$$S_{shared} = S_{shared,obs} + f_{0+} + f_{+0} + f_{00}. \tag{6}$$

The four terms in the right hand side of Eq. (6), as defined earlier, represent, respectively, the number of shared species observed in both samples (f_{++}), the number of shared species observed only in Sample II (f_{0+}), the number of shared species observed only in Sample I (f_{+0}), and the number of shared species undetected in both samples (f_{00}), but present in both assemblages. Only the first term is observable. The sum of the last three terms represents the total undetected shared species richness. Applying Eqs. (5a) and (5b) and using notation and derivation similar to Eq. (2c), we have

$$\hat{f}_{0+} = \frac{\widehat{\alpha_{0+} f_{0+}}}{\hat{\alpha}_{0+}} \geq \frac{\widehat{\alpha_{0+} f_{0+}}}{\hat{\alpha}_{1+}} = \frac{\frac{f_{1+}}{n_1}}{\frac{2f_{2+}}{(n_1-1)f_{1+}}} = \frac{(n_1 - 1)}{n_1} \frac{f_{1+}^2}{2f_{2+}}. \tag{7a}$$

and a symmetric expression

$$\hat{f}_{+0} = \frac{\widehat{\alpha_{+0} f_{+0}}}{\hat{\alpha}_{+0}} \geq \frac{\widehat{\alpha_{+0} f_{+0}}}{\hat{\alpha}_{+1}} = \frac{\frac{f_{+1}}{n_2}}{\frac{2f_{+2}}{(n_2-1)f_{+1}}} = \frac{(n_2 - 1)}{n_2} \frac{f_{+1}^2}{2f_{+2}}. \tag{7b}$$

Furthermore, we expect that $\alpha_{00} \leq \alpha_{11}$, and this ordering is preserved by the corresponding estimates, implying the following inequality, from Eq. (5c):

$$\begin{aligned} \hat{f}_{00} = \frac{\widehat{\alpha_{00} f_{00}}}{\hat{\alpha}_{00}} \geq \frac{\widehat{\alpha_{00} f_{00}}}{\hat{\alpha}_{11}} &= \frac{\frac{f_{11}}{n_1 n_2}}{\frac{4f_{22}}{(n_1-1)(n_2-1)f_{11}}} \\ &= \frac{(n_1 - 1)}{n_1} \frac{(n_2 - 1)}{n_2} \frac{f_{11}^2}{4f_{22}}. \end{aligned} \tag{7c}$$

Combining (6) and (7a) – (7c), we obtain the following estimator of shared species richness:

$$\begin{aligned} \hat{S}_{Chao1,shared} = S_{shared,obs} &+ k_1 \frac{{f_{1+}}^2}{2f_{2+}} + k_2 \frac{{f_{+1}}^2}{2f_{+2}} \\ &+ k_1 k_2 \frac{{f_{11}}^2}{4f_{22}}, \end{aligned} \tag{7d}$$

where $k_i = (n_i - 1)/n_i$, $i = 1, 2$. A modification similar to that in Eq. (3a) can be applied to each term to avoid a zero divisor. The sum of the last three terms estimates the undetected shared species richness. The resulting estimator in Eq. (7d) is identical to that derived by Pan et al. (2009), who derived it by means of complicated mathematical inequalities. Here we show that Good-Turing's framework provides a simple, unified approach to inferring undetected diversity, not only for one assemblage but also for two assemblages. The estimator in Eq. (7d) is referred to as the *Chao1-shared* estimator because it can be regarded as an extension of the single-assemblage Chao1 estimator (Eq. 3a) to the case of two assemblages. Pan et al. (2009) also derived a variance estimator by using a standard approximation theory, allowing construction of a confidence interval for true shared species richness. From the above derivation, the Chao1-shared lower bound becomes a nearly unbiased point estimator provided $\hat{\alpha}_{0+} \approx \hat{\alpha}_{1+}, \hat{\alpha}_{+0} \approx \hat{\alpha}_{+1}$, and $\hat{\alpha}_{00} \approx \hat{\alpha}_{11}$ in the derivations. Considering only shared species, we see that a simple sufficient condition is that the undetected species and the detected singletons have approximately the same abundances in each of the two assemblages.

Phylogenetic Diversity

Undetected Faith's PD in a single assemblage

To formulate phylogenetic diversity, we assume that all S species in an assemblage are connected by a rooted ultrametric or non-ultrametric phylogenetic tree, with all species, observed and unobserved, as tip nodes. In this paper, all phylogenetic diversity measures and estimators are computed from a given fixed reference point that is ancestral to all taxa considered in the study. The choice of the reference point is thus independent of the sampling data. Assume that there are B branch segments and B corresponding nodes, $B \geq S$. Let a_i denote the total relative abundance of the species descended from the ith node/branch, $i = 1, 2, \ldots, B$, and L_i denote the length of branch i. Therefore, the set of species relative abundances $(p_1, p_2, \ldots, p_S)$ is expanded to a larger relative abundance set $\{a_i, i = 1,\ 2,\ \ldots, B\}$ with $(p_1, p_2, \ldots, p_S)$ as its first S elements. For simplicity, we refer to a_i as the *node/branch relative abundance* of the ith node/branch, although $\sum_{i=1}^{B} a_i$ is not necessarily equal to unity; see Fig. 1 of Chao et al. (2015*a*) for an illustrative example. Faith's (1992) PD is expressed as $PD = \sum_{i=1}^{B} L_i$.

We assume an empirical sample of n individuals with sample species abundances $(X_1,\ X_2,\ \ldots, X_S)$ is taken from the assemblage. Define X_i^* as the sum of the observed species abundances for those species in the

sample that are descended from branch i. Then we can expand the set of observed species abundances to a larger branch abundance set $\{X_i^*,\ i = 1,\ 2,\ \ldots,\ B\}$ with $(X_1,\ X_2,\ \ldots, X_S)$ as its first S elements. We refer to X_i^*, $i = 1,\ 2,\ \ldots,\ B$, as the *sample node/branch abundance* of node/branch i. Let

$$g_r = \sum\nolimits_{i=1}^{B} L_i\, I(X_i^* = r), r = 0, 1, 2, \ldots, \qquad (8a)$$

be the total length of those branches with sample abundance r in the set $\{X_i^*,\ i = 1,\ 2,\ \ldots,\ B\}$, where the indicator function $I(\cdot)$ is defined in Eq. (1a). The undetected PD in the sample is g_0, which is the total length of undetected branches; g_0 is unknown but $\{g_1,\ g_2,\ \ldots\}$ can be computed from the sample and the observed tree (the tree spanned by the observed species). For $r > 0$, g_r is identical to the total length of the branches with sample abundance r in the observed tree. For example, g_1 denotes the total length of those nodes/branches with sample abundance = 1 in the observed tree; g_2 denotes the total length of those branches with sample abundance = 2 in the observed tree. Let PD_{obs} denote the observed PD. Then we have $PD_{obs} = \sum_{i>0} g_i \equiv g_+$ and $PD = PD_{obs} + g_0$. A simple example is provided in Appendix S2: Fig. S1 to illustrate these measures.

We now extend the Good-Turing frequency formula to its phylogenetic version. For those nodes/branches that each is with a sample abundance/frequency of r, the branch-length-weighted mean (per unit length) of their true abundances in the entire assemblage, λ_r, can be mathematically expressed as

$$\lambda_r = \sum\nolimits_{i=1}^{B} L_i a_i\, I(X_i^* = r)/g_r, r = 0, 1, 2, \ldots$$

Derivation steps parallel to those used for the species diversity lead to the following phylogenetic version of the single-assemblage Good-Turing frequency formula (Appendix S2):

$$\hat{\lambda}_r = \frac{(r+1)g_{r+1}}{(n-r)g_r}, r = 0, 1, 2\ldots \qquad (8b)$$

Intuitively, we expect that $\lambda_0 \le \lambda_1$, and this ordering is preserved by the corresponding estimates. From the above formula, a lower bound for the undetected PD is obtained:

$$\hat{g}_0 = \frac{\widehat{\lambda_0 g_0}}{\hat{\lambda}_0} \ge \frac{\widehat{\lambda_0 g_0}}{\hat{\lambda}_1} = \frac{\frac{g_1}{n}}{\frac{2g_2}{(n-1)g_1}} = \frac{(n-1)}{n}\frac{g_1^2}{2g_2}. \qquad (8c)$$

The lower bound in Eq. (8c) is identical to that proposed in Chao et al. (2015*a*) via the Cauchy-Schwarz inequality. Here, instead, we prove it under a unified framework by means of a phylogenetic version of the original Good-Turing formula. Cardoso et al. (2014) adapted the Chao1 estimator (Eq. 3a) to obtain the same estimator.

When g_2 is relatively small, including the case of $g_2 = 0$, the above estimator may yield an extremely large value and thus exhibit a large variance. To cope with such cases, Chao et al. (2015*a*) and Hsieh and Chao (2017) proposed the following modified Chao1-PD estimator:

$$\widehat{PD}_{Chao1} = \begin{cases} PD_{obs} + \dfrac{(n-1)}{n}\dfrac{g_1^2}{2g_2}, & \text{if } g_2 > \dfrac{g_1 f_2^*}{2f_1^*}; \\ PD_{obs} + \dfrac{(n-1)g_1(f_1^*-1)}{n\quad 2(f_2^*+1)}, & \text{if } g_2 \le \dfrac{g_1 f_2^*}{2f_1^*}, \end{cases} \qquad (8d)$$

where f_1^* and f_2^* denote, respectively, the number of nodes/branches with abundance = 1 and abundance = 2 in the observed tree. Since any node/branch with abundance = 1 can occur only at the tip nodes, we have $f_1^* = f_1$ i.e., f_1^* is identical to the number of singletons at the tip nodes. However, f_2^* is not necessarily equal to f_2; see Appendix S2: Fig. S1 for an example.

Based on Eq. (8c), the Chao1-PD estimator is nearly unbiased when $\hat{\lambda}_1 \approx \hat{\lambda}_0$, which means that the singletons (which occur only at the tip nodes) and the undetected nodes (which can be tip or interior nodes) have the same length-weighted mean abundances. A sufficient simple condition is that all rare nodes (including the singletons and the undetected) have approximately the same abundances. The variance of the Chao1-PD estimator can be obtained using Eq. (3b) with $\{f_1, f_2\}$ being replaced by $\{g_1, g_2\}$. The construction of the confidence interval for Faith's PD based on the Chao1-PD estimator can be similarly obtained.

From the derivations above, we see that all estimation procedures and the derivation steps for developing the phylogenetic version of the Good-Turing formula and PD estimators are parallel to those for the original formula and species richness estimators. A summary of formulas and descriptions for estimating species richness and Faith's PD is provided in Appendix S2: Table S1, where the analogy between the two estimation frameworks is transparently displayed. The analogy was first proposed by Faith (1992). From Faith's perspective, each unit-length branch is regarded as a "feature" in phylogenetic diversity (like a "species" in species diversity). Chao et al. (2014*a*) subsequently referred to each unit-length branch segment as a *phylogenetic entity*. For example, a branch of 8-unit length is counted as 8 phylogenetic entities. All entities are phylogenetically equally distinct, just as all species are assumed taxonomically equally distinct in computing species richness. Instead of species, for PD we are measuring the total number of phylogenetic entities, or equivalently, the total branch length (because each entity has length of unity). Using this perspective, the measures of branch lengths $\{g_k,\ k = 0,\ 1,\ \ldots,\ \}$ in estimating PD play the same role as the frequency counts $\{f_k,\ k = 0,\ 1,\ \ldots\}$ in estimating species richness. This analogy to counting-up species means that most ecological indices defined at the species level can be converted to PD equivalents (by counting phylogenetic entities rather than species).

Undetected shared PD between two assemblages

Following the approach to the two-assemblage model formulation and the data framework described in the section *Two-assemblage Good-Turing Formulas*, we assume that all S species of the *pooled* assemblage are indexed by 1, 2, …, S. Assume these S species are connected by a rooted ultrametric or non-ultrametric phylogenetic tree, with the S species as tip nodes. Given a fixed reference point that is ancestral to all taxa considered in the study, we assume that there are B branch segments and B corresponding nodes in the pooled tree, and let L_i denote the length of branch i. Because the phylogenetic tree of each individual assemblage is a subtree of the pooled tree, the diversity for each individual assemblage can be computed from the pooled tree structure with only the node (or branch) abundances varying between assemblages, as illustrated by Chiu et al. (2014*a*, Figure 2). Assume that there are B_{12} branches shared by the two assemblages for the given reference point, and without losing generality, these shared branches are indexed from 1, 2, …, B_{12} for notational simplicity. Denote the true shared *PD* (the total length of branch segments shared by the two individual subtrees) by PD_{shared} and the shared *PD* between two observed trees by $PD_{shared,obs}$.

As in the preceding section, the two sets of relative species abundances in Assemblages I and II are, respectively, expanded to the *node/branch relative abundance (or frequency)* sets $\{a_{i1}, i = 1, 2, \ldots, B\}$ and $\{a_{i2}, i = 1, 2, \ldots, B\}$. We also expand the set of observed species frequencies $(X_{11}, X_{21}, \ldots, X_{S1})$ in Assemblage I to a larger sample node/branch frequency set $\{X^*_{i1}, i = 1, 2, \ldots, B\}$. Similarly, we extend the set of observed species frequencies $(X_{12}, X_{22}, \ldots, X_{S2})$ in Assemblage II to a larger set $\{X^*_{i2}, i = 1, 2, \ldots, B\}$. The frequency counts of shared species, f_{rv}, (Eq. 4a), are also extended to their phylogenetic versions:

$$g_{rv} = \sum_{i=1}^{B_{12}} L_i\, I(X^*_{i1} = r,\ X^*_{i2} = v),\ r, v = 0, 1, 2, \ldots \quad (9)$$

Here g_{rv} measures the total length of those shared branches with node/branch abundance r in Sample I and node/branch abundance v in Sample II. We can similarly define g_{r+}, g_{+v} and g_{++}, as we did in Eqs. (4b)–(4d). Here g_{r+} denotes the total length of those shared branches that have abundance r in Sample I and non-zero node/branch abundance in Sample II, with a similar interpretation for g_{+v}. Now g_{++} becomes the observed *PD* shared by the two samples, i.e., $g_{++} = PD_{shared,obs}$. Thus $\{g_{rv}, g_{r+}, g_{+v}, g_{++}; r, v = 0, 1, 2\ldots\}$ are analogous to $\{f_{rv}, f_{r+}, f_{+v}, f_{++}; r, v = 0, 1, 2\ldots\}$ as defined in Eqs. (4a) – (4d). The only difference is that each of the former set measures the lengths of shared branches, whereas each of the later set counts shared species.

The true shared *PD* can be expressed as the sum of four terms:

$$PD_{shared} = PD_{shared,obs} + g_{0+} + g_{+0} + g_{00}. \quad (10a)$$

The four terms in the right hand side of the above equation represent, respectively, the observed shared *PD*, the shared *PD* that is missed only in Sample I, the shared *PD* that is missed only in Sample II, and the shared *PD* that is missed by both samples. For any shared branch segment with node/branch abundance r in each observed branch set, consider calculating the product of its true branch relative abundances in the two assemblages. The branch-length-weighted mean (per unit-length) of the products among all such shared branches (their total length is g_{rr}) can be expressed as

$$\lambda_{rr} = \sum_{i=1}^{B_{12}} L_i a_{i1} a_{i2}\, I(X^*_{i1} = r,\ X^*_{i2} = r)/g_{rr}, r = 0, 1, 2\ldots$$

The two-assemblage phylogenetic version of the Good-Turing formula (see Appendix S2 for a proof) provides the following estimator of λ_{rr} which is similar to that for shared species richness in Eq. (5c):

$$\hat{\lambda}_{rr} = \frac{(r+1)^2 g_{r+1,r+1}}{(n_1 - r)(n_2 - r) g_{rr}}, r = 0, 1, 2, \ldots \quad (10b)$$

Under the intuitive expectation that $\lambda_{00} \le \lambda_{11}$ (and this ordering is preserved by the corresponding estimates), we obtain the following lower bound of g_{00}:

$$\hat{g}_{00} = \frac{\widehat{\lambda_{00} g_{00}}}{\hat{\lambda}_{00}} \ge \frac{\widehat{\lambda_{00} g_{00}}}{\hat{\lambda}_{11}} = \frac{\frac{g_{11}}{n_1 n_2}}{\frac{4g_{22}}{(n_1-1)(n_2-1)g_{11}}} = \frac{(n_1 - 1)}{n_1}\frac{(n_2 - 1)}{n_2}\frac{g_{11}^2}{4g_{22}}. \quad (10c)$$

Then the same estimation procedures we used to develop the Chao1-shared estimator in Eq. (7d) lead to the following Chao1-*PD*-shared estimator (Appendix S2):

$$\widehat{PD}_{Chao1} = PD_{shared,obs} + k_1 \frac{g_{1+}^2}{2g_{2+}} + k_2 \frac{g_{+1}^2}{2g_{+2}} + k_1 k_2 \frac{g_{11}^2}{4g_{22}}, \quad (10d)$$

where $k_i = (n_i - 1)/n_i$, $i = 1, 2$. The sum of the last three terms estimates the total length of undetected branches shared by two individual sub-trees. A modification similar to that proposed in Eq. (8d) can be applied to each of the three terms. The sufficient conditions for the Chao1-*PD*-shared estimator being nearly unbiased can be similarly formulated as those for the Chao1-shared estimator, simply by replacing "shared species" with "shared nodes/branches."

The variance and confidence interval associated with this estimator follow directly from those for the Chao1-

shared estimator, by replacing the counts $\{f_{rv}, f_{r+}, f_{+v}, f_{++}; r, v = 0, 1, 2, \ldots\}$ with $\{g_{rv}, g_{r+}, g_{+v}, g_{++}; r, v = 0, 1, 2, \ldots\}$ in the formulas. This correspondence also reflects the analogy between the estimation of shared species and shared *PD* between two assemblages. A comparison of estimation formulas used in estimating shared species richness and in estimating shared *PD* is provided in Appendix S2: Table S2, where any unit-length branch shared by two assemblages plays the same role as a "shared species" in the estimation of shared species richness. The measures of shared branch lengths $\{g_{rv};\ r, v = 0,\ 1,\ 2,\ \ldots\}$ used in estimating shared *PD* play a role corresponding to the frequency counts $\{f_{rv};\ r, v = 0,\ 1,\ 2,\ \ldots\}$ in estimating shared species diversity. Hence, not only within-assemblage but also between-assemblage measures defined at the species level can be converted to *PD* equivalents.

Functional Diversity

Undetected FAD in a single assemblage

Consider an assemblage in which all species are characterized by a set of functional traits, which can be categorical or continuous variables. The pairwise distances are calculated by some distance metric (e.g., Euclidean distance or Gower distance) based on the values of species traits. Let d_{ij} be the functional distance between the ith and jth species, with $d_{ij} = d_{ji} > 0$. Walker et al.'s (1999) *FAD* in the entire assemblage is expressed as $FAD = \sum_{i,j=1}^{S} d_{ij}$. In our estimation procedures, we do not require zero distance between any two species. That is, intraspecific variability in traits is allowed, so that d_{ii} may be greater than zero. For example, when trait values are available at the individual level, we can simply define the distance between two species as the mean of all distances between any pair of individuals, one chosen respectively from each of the two species.

In our approach, one unit length of distance is analogous to a pair of species shared between two identical assemblages. We first extend all frequency counts of shared species, f_{rv}, defined in Eqs. (4a)–(4d) to the corresponding measures of pairwise distance based on a sample of n individuals, with a species sample abundance set $(X_1,\ X_2,\ \ldots, X_S)$. Let F_{rv} denote the total distance of those species pairs with sample abundances r and v respectively for the two species in each pair. That is,

$$F_{rv} = \sum_{i,j=1}^{S} d_{ij}\ I(X_i = r, X_j = v). \tag{11}$$

We can similarly define F_{r+}, F_{+v} and F_{++}, as we did in Eqs. (4b) – (4d). Here F_{r+} denotes the total distance of those species pairs with sample abundance r and k, where $k > 0$ can be any positive integer, with a similar definition for F_{+v}, while F_{++} becomes the observed *FAD* in the sample, i.e., $F_{++} = FAD_{obs}$.

Given one-sample data (data for a single assemblage) and the species-pairwise distance matrix for all species, the true *FAD* can be expressed as the sum of four terms:

$$FAD = FAD_{obs} + F_{0+} + F_{+0} + F_{00}. \tag{12a}$$

The four terms in the right hand side of the above equation denote, respectively, the observed *FAD*, the undetected *FAD* for species pairs in which one species is missing from the sample ($F_{0+} = F_{+0}$), and the undetected *FAD* for species pairs in which both species are missing from the sample (F_{00}). This expression is analogous to Eq. (6) for shared species and to Eq. (10a) for shared *PD*. Out of all the species which appeared exactly r times in the sample, consider choosing any two to form a pair and calculating the product of their true relative abundances in the assemblage. The distance-weighted mean (per unit-distance) of the products among all such pairs (their total pairwise distance is F_{rr}) can be expressed as

$$\theta_{rr} = \sum\nolimits_{i,j=1}^{S} d_{ij}\, p_i\, p_j\, I(X_i = r,\ X_j = r)/F_{rr}, r = 0, 1, 2\ldots$$

The functional version of the Good-Turing frequency formula provides the following estimator of θ_{rr} (see Appendix S3 for a proof):

$$\hat{\theta}_{rr} = \frac{(r+1)^2 F_{r+1,r+1}}{(n-2r)(n-2r-1)F_{rr}}, r = 0, 1, 2, \ldots \tag{12b}$$

The above formula is slightly different from the two-sample Good-Turing formulas for estimating shared species (Eq. 5c) and shared *PD* (Eq. 10b), because, here, we only have one-sample data taken from a single assemblage, whereas the data for the previous formulas are taken from two different assemblages. Under the intuitive expectation that $\theta_{00} \le \theta_{11}$ (and this ordering is preserved by the corresponding estimates), we obtain the following lower bound for F_{00}:

$$\begin{aligned}\hat{F}_{00} = \frac{\widehat{\theta_{00}F_{00}}}{\hat{\theta}_{00}} \ge \frac{\widehat{\theta_{00}F_{00}}}{\hat{\theta}_{11}} &= \frac{\frac{F_{11}}{n(n-1)}}{\frac{4F_{22}}{(n-2)(n-3)F_{11}}} \\ &= \frac{(n-2)}{n}\frac{(n-3)}{n-1}\frac{F_{11}^2}{4F_{22}}.\end{aligned} \tag{12c}$$

Then the same estimation procedures as we obtained for Chao1-shared in Eq. (7d) leads to the following Chao1-*FAD* estimator:

$$\begin{aligned}\widehat{FAD}_{Chao1} =& FAD_{obs} + k\frac{F_{1+}^2}{2F_{2+}} + k\frac{F_{+1}^2}{2F_{+2}} \\ &+ \frac{(n-2)(n-3)}{n(n-1)}\frac{F_{11}^2}{4F_{22}},\end{aligned} \tag{12d}$$

where $k = (n-1)/n$. The sum of the last three terms estimates the undetected *FAD*. A modification similar to

that proposed in Eq. (8d) can be applied to each of the three terms. The variance and confidence interval associated with this estimator follow directly from those for the Chao1-shared estimator. Under the condition that undetected species and singletons have approximately homogenous abundances, the Chao1-*FAD* estimator is nearly unbiased for any given species-pairwise distance matrix.

A summary of formulas and descriptions for estimating shared species richness and *FAD* is given in Appendix S3: Table S1, where the analogy between the estimation procedures of the two measures can be seen. Chao et al. (2014*a*) define a "functional entity" as a species pair with one unit of distance between the two species. In *FAD*, a functional entity plays the same role as a "shared species" between two assemblages. For example, a species-pair with distance $d_{ij} = 5$ is counted as 5 "shared species" (i.e., 5 functional entities). Thus the measures of total distances of species pairs, $\{F_{rv}, F_{r+}, F_{+v}, F_{++}; r, v = 0, 1, 2, \ldots\}$, play the same roles as the counts of shared species richness $\{f_{rv}, f_{r+}, f_{+v}, f_{++}; r, v = 0, 1, 2, \ldots\}$ (defined in Eqs. 4a – 4d).

Undetected shared FAD between two assemblages

Under the two-assemblage model formulation and data framework described in the section *Two-assemblage Good-Turing Formulas*, we further assume that the functional distance between the *i*th and *j*th species for all *S* species of the *pooled* assemblage is denoted as d_{ij}, with $d_{ij} = d_{ji} > 0$. The *FAD* of each individual assemblage can be computed from a sub-matrix of the distance matrix of all *S* species in the pooled assemblage. Denote the number of shared species by S_{12}, shared *FAD* (i.e., the total species-pairwise distance of two species each is shared by two assemblages) by FAD_{shared}, and the *FAD* based on the observed shared species by $FAD_{shared,obs}$.

Given two-sample data with species sample abundances $(X_{11}, X_{21}, \ldots, X_{S1})$ and $(X_{12}, X_{22}, \ldots, X_{S2})$, as defined earlier, we extend all frequency counts F_{rv} defined in Eq. (11) to the two-assemblage case. That is, we define for $r, v, k, m = 0, 1, 2\ldots$

$$F_{(rv)(km)} = \sum_{i,j=1}^{S_{12}} d_{ij}\, I(X_{i1} = r,\ X_{j1} = v) \times I(X_{i2} = k,\ X_{j2} = m). \tag{13}$$

Here $F_{(rv)(km)}$ measures the total functional distance of those pairs of shared species with abundances (r, v) in Sample I and abundances (k, m) in Sample II for the two species in each pair. Where any index is replaced by a "+" sign, it means that those species occur at least once in the corresponding sample. Thus, we can define $F_{(+v)(km)}$, $F_{(+v)(+m)}$, $F_{(++)(km)}$, and $F_{(rv)(+m)}$, etc. Then shared *FAD* can be expressed as the sum of 16 terms (see Appendix S3: Table S2 for illustrative details)

$$\begin{aligned} FAD_{shared} &= \sum_{r,v=+,0} \sum_{k,m=+,0} F_{(rv)(km)} \\ &= F_{(++)(++)} + F_{(++)(+0)} + F_{(++)(0+)} \\ &\quad + \ldots + F_{(00)(00)}. \end{aligned} \tag{14a}$$

Here $F_{(++)(++)} = FAD_{shared,obs}$ denotes the observed shared *FAD*; the other 15 terms are unknown. For each term, we can apply similar-type Good-Turing formulas to obtain an estimator as given in Appendix S3: Table S3. Then we have the following Chao1-*FAD*-shared estimator:

$$\begin{aligned} \widehat{FAD}_{Chao1,shared} &= F_{(++)(++)} + \hat{F}_{(++)(+0)} \\ &\quad + \hat{F}_{(++)(0+)} + \ldots + \hat{F}_{(00)(00)}. \end{aligned} \tag{14b}$$

A bootstrap method can be used to assess the sampling variance of the above estimator and to obtain the associated confidence interval.

Example

We apply all the estimators derived from the Good-Turing frequency formula and its generalizations to the rain-forest tree data described and discussed in Magnago et al. (2014). The tree species abundance data were collected between January 2011 and January 2012 from 11 forest fragments in Espírito Santo State, in southeastern Brazil. Sampling data for 10 fragments included one Edge and one Interior transect, whereas in one fragment sampling data included two Edge and two Interior transects (Magnago et al. 2014), with a distance of 5.7 ± 2.4 km between transects. In total, the study comprised 11 fragments and 24 transects. Each Edge transect was placed about 5 m inside the fragment and parallel to the forest edge, and each Interior transect was located at least 300 m from the nearest edge. One of the goals of the original study was to compare functional diversity between Edge habitat and Interior habitat. Within each transect, every living tree with a diameter at breast height (DBH) > 4.8 cm and 1.3 m height was recorded.

In the original data (Table S2 of Magnago et al. 2014), pooled from all 24 transect samples in 11 fragments, there were 443 species among 4140 individual trees. However, in order to construct the phylogenetic tree for the observed species, using Phylomatic (http://www.phylodiversity.net/phylomatic; Webb and Donoghue 2005), we had to exclude 18 species, including 5 species not identified in the species list of Magnago et al. (2014) and 13 species not included in Phylomatic. Among the 18 species excluded, 12 species were found in both habitats, 3 species were unique to the Edge habitat and 3 species were unique to the Interior habitat. The data considered in our analysis thus include 425 species from a total of 3868 individuals; species abundance data are available in Github (https://github.com/AnneChao). The species abundance frequency counts of these 425 species are summarized in Table 1. There were 319

TABLE 1. The species frequency counts of the tree abundance data from two habitats (Edge and Interior) in south-eastern Brazil (Magnago et al. 2014), where f_i denotes the number of species represented by exactly i individuals in the sample.

Number of individuals in sample, i	Number of species, f_i	
	Edge	Interior
1	110	123
2	48	48
3	38	41
4	28	32
5	13	19
6	11	17
7	12	6
8	5	7
9	4	6
10	6	7
11	3	6
12	4	3
13	3	3
14	5	3
15	2	3
16	4	4
17	2	2
18	2	1
19	2	3
20	2	4
21	1	5
23	2	2
25	1	1
27	1	2
28	1	1
30		1
32	1	2
34		1
35		1
36	2	
37	1	
41	1	
45	1	
46	1	
49	1	
52		1
110	1	
123		1
140		1

Notes: Detailed species abundances are taken from Table S2 of Magnago et al. (2014). For the Edge habitat, $S_{obs} = 319$, $n = 1794$, sample coverage estimate = 93.9%; for the Interior habitat, $S_{obs} = 356$, $n = 2074$, sample coverage estimate = 94.1%.

species (including 110 singletons and 48 doubletons) among 1794 individuals in the data from the Edge habitat, and 356 species (including 123 singletons and 48 doubletons) among 2074 individuals in the data from the Interior habitat. There were 250 species in common between the two habitats in the data. The sample coverage estimates for these two habitats are nearly equal (93.9% for the Edge habitat and 94.1% for the Interior habitat), in spite of different sample sizes.

To illustrate our estimators, we focus only on the diversity analysis for the pooled samples over all fragments. That is, we regard the pooled tree records as sampling data from the entire study area represented by the 11 fragments and aim to infer and compare the estimated true or asymptotic diversity (observed plus undetected) of the two habitats for the whole study area. We could also apply our estimation to the Edge and Interior transect data within each fragment to compare the diversity of the two habitats; in this case, the target becomes the estimated true diversity of each fragment. Comparison of compositional, phylogenetic, and functional differentiation among the 11 fragments, i.e., beta diversity with adjustment for under-sampling bias will be reported elsewhere.

Estimates for species richness and shared species richness between the two habitats appear in Table 2. The undetected species richness estimates (Eq. 2c) for the Edge and Interior habitats are, respectively, 126 and 158 species, implying that the total richness estimate (Eq. 3a) for the Edge habitat is 445 species, with a 95% confidence interval of (396, 525), and for the Interior habitat 514 species, with a 95% confidence interval of (455, 609). These results show that the Interior habitat has higher estimated species richness; however, the difference is not statistically significant at a level of 5%. The undetected shared species richness between the two habitats is estimated to be 139, leading to an estimate of shared richness (Eq. 7d) of 389 species, with a 95% confidence interval of (347, 450). Although the true species richness and shared species richness between the two habitats are unknown and our estimates theoretically represent lower bounds, the data collector among us (L. F. S. Magnago) believes these estimates provide reasonable adjustments based on his experiences in the fields, CVRD (Herbarium of Natural Vale Reserve) herbarium collection, and other floristic studies for the region (Rolim and Nascimento 1997, Jesus and Rolim 2005, Paula and Soares 2011). We also assume that this conclusion can be extended to the following estimates of phylogenetic diversity and functional diversity. The biological processes that may give rise detection failure are elaborated in the *Discussion* section.

The phylogenetic tree (in Newick format) of the 425 observed species which was constructed using the software Phylomatic (Webb and Donoghue 2005) is given in Github (https://github.com/AnneChao). The observed Faith's *PD* values are, respectively, 24516 and 27727 Myr in the Edge and Interior habitat, with a shared *PD* 20680 Myr. Table 3 shows the estimates for Faith's *PD* in each habitat and shared *PD* between the two habitats. The *PD* estimates (Eq. 8d) for the Edge and Interior habitats are, respectively, 32011 with a 95% confidence interval of (31542, 32511), and 34550 with a 95% confidence interval of (34143, 34983). The two intervals do not overlap, implying that the Interior habitat has significantly higher *PD* than the Edge habitat. The shared *PD* between the two habitats is estimated (Eq. 10d) to be 29360 Myr, with a 95% confidence interval of (28976, 29761).

TABLE 2. Summary of species data and richness estimates based on tree species abundance data for the Edge and Interior habitats in forest fragments of south-eastern Brazil (Magnago et al. 2014), showing (a) undetected species richness and Chao1 richness point and interval estimates for each habitat (see Eq. 3a) and (b) undetected shared species richness between the two habitats and the corresponding Chao1-shared richness point and interval estimates (see Eq. 7d).

Habitat	Sample size	f_1	f_2	Observed richness	Undetected richness	Chao1 richness	95% conf. interval
(a)							
Edge	1794	110	48	319	126	445	396, 525
Interior	2074	123	48	356	158	514	455, 609

Observed shared richness	f_{+1}	f_{+2}	f_{1+}	f_{2+}	f_{11}	f_{22}	$\hat{f}_{+0}$	$\hat{f}_{0+}$	$\hat{f}_{00}$	Undetected shared richness	Chao1-shared richness	95% conf. interval
(b)												
250	64	30	60	37	25	7	68	49	22	139	389	347, 450

The functional traits for these tree data were based primarily on morphological and physical characteristics of trees, their roles as trophic resources, their dispersal modes, and their roles in carbon storage and forest structure. All observed species were described by a set of six functional traits, including five categorical variables: fruit size (size categories), seed size (size categories), fruit type, fruit dispersal syndrome, and successional group, together with one quantitative variable: wood density. Based on these six traits, the species pairwise distance matrix for the data (pooled across replicates) was calculated by a Gower mixed-variables coefficient of distance with equal weights for all traits; the Gower distance matrix of the 425 observed species is provided in Github (https://github.com/AnneChao). Table 4 shows the estimates for *FAD* in each habitat and shared *FAD* between the two habitats. In the Edge and Interior habitats, the observed *FAD* (total pairwise Gower distances) are 36603 and 43438, respectively. Based on Eq. (12d), the *FAD* estimate is 69670 in the Edge habitat, with a 95% confidence interval of (68764, 70602), and 86802 in the Interior habitat, with a 95% confidence interval of (85659, 87975). Here each estimate of true *FAD* is almost double the observed *FAD*. We can conclude that the *FAD* in the Interior habitat is significantly higher than the *FAD* in the Edge habitat because the two intervals do not overlap. The shared *FAD* between the two habitats is estimated (Eq. 14b) to be 49061 with a 95% bootstrap confidence interval of (42034, 58562) based on 200 bootstrap replications. This wide confidence interval reflects large sampling variance due to the estimation of 15 parameters and each estimate is subject to some degree of sampling uncertainty.

Our analysis suggests that the pooled Interior habitat has significantly higher phylogenetic diversity and significantly higher functional diversity than the pooled Edge habitat, but no such significance can be concluded for species diversity. These results are generally consistent with previous findings (Magnago et al. 2014, Matos et al. 2017) based on empirical data and model expectations. The previous authors found that the Edge habitat characteristics (microclimate, light, etc.) do not change with increasing fragment size whereas the Interior habitat characteristics change with fragment size. Consequently, fragment size usually does not produce effects on species richness and composition at fragment edges, but strongly influences the diversity in the interior of fragments. Because all fragment sizes are considered together in this paper, we thus expect greater diversity values for Interior than for Edge habitats.

TABLE 3. Summary of phylogenetic data and diversity estimates for the Edge and Interior habitats in forest fragments of south-eastern Brazil (Magnago et al. 2014), showing (a) undetected *PD* and Chao1-*PD* point and interval estimates for each habitat (see Eq. 8d), and (b) undetected shared *PD* between the two habitats and the corresponding Chao1-*PD*-shared point and interval estimates (see Eq. 10d).

Habitat	Sample size	g_1	g_2	Observed *PD*	Undetected *PD*	Chao1-*PD*	95% conf. interval
(a)							
Edge	1794	6578	2885	24516	7495	32011	31542, 32511
Interior	2074	7065	3656	27727	6823	34550	34143, 34983

Observed shared *PD*	g_{+1}	g_{+2}	g_{1+}	g_{2+}	g_{11}	g_{22}	$\hat{g}_{+0}$	$\hat{g}_{0+}$	$\hat{g}_{00}$	Undetected shared *PD*	Chao1-*PD* shared	95% conf. interval
(b)												
20680	3888	2177	3929	2125	1711	463	3470	3630	1579	8680	29360	28976, 29761

Note: The phylogenetic tree for all observed species is based on Phylomatic (Webb and Donoghue 2005).

TABLE 4. Summary of functional data and diversity estimates based on tree species abundance data, as well as species traits collected from the Edge and Interior habitats of forest fragments in south-eastern Brazil (Magnago et al. 2014), showing (a) undetected *FAD* and Chao1-*FAD* point and interval estimates for each habitat (see Eq. 12d), and (b) undetected shared *FAD* between the two habitats and the corresponding Chao1-*FAD*-shared point and interval estimates (see Eq. 14b and Appendix S3).

Habitat	Observed *FAD*	$F_{+1}(=F_{1+})$	$F_{+2}(=F_{2+})$	F_{11}	F_{22}	$\hat{F}_{+0}=\hat{F}_{0+}$	$\hat{F}_{00}$	Undetected *FAD*	Chao1-*FAD*	95% conf. interval
(a)										
Edge	36603	12452	5572	4200	837	13906	5256	33068	69670	68764, 70602
Interior	43438	14769	6059	4940	825	17992	7380	43364	86802	85659, 87975

Observed shared *FAD*	$\hat{F}_{(++)(00)}$	$\hat{F}_{(++)(0+)}=\hat{F}_{(++)(+0)}$	$\hat{F}_{(00)(++)}$	$\hat{F}_{(+0)(++)}=\hat{F}_{(0+)(++)}$	$\hat{F}_{(+0)(0+)}=\hat{F}_{(0+)(+0)}$	$\hat{F}_{(+0)(00)}=\hat{F}_{(0+)(00)}$	$\hat{F}_{(+0)(+0)}=\hat{F}_{(0+)(0+)}$	$\hat{F}_{(00)(+0)}=\hat{F}_{(00)(0+)}$	$\hat{F}_{(00)(00)}$	Undetected shared *FAD*	Chao1-*FAD* shared	95% conf. interval
(b)												
22079	727	5793	671	3906	1028	166	1753	128	34	26981	49061	42034, 58562

Note: The functional distance matrix between any two observed species in the pooled assemblage is obtained by Gower distance.

Conclusion and Discussion

We have generalized the original one-assemblage Good-Turing frequency formula (Eqs. 1c and 1d) to the case of two assemblages (Eq. 5c), and also extended it to a phylogenetic version (Eqs. 8b and 10b) as well as a functional trait version (Eq. 12b and Appendix S3: Table S3). We have also applied the original and generalized formulas to obtain various estimators of undetected species, phylogenetic, and functional diversity, as summarized below.

1. For species diversity, the estimator of species richness derived from the Good-Turing frequency formula is identical to the Chao1 estimator (Chao 1984, 1987) in Eq. (3a). The estimator of shared species between two assemblages in Eq. (7d) is identical to the Chao1-shared species estimator proposed in Pan et al. (2009).
2. For phylogenetic diversity, the resulting estimator (Eq. 8d) of Faith's *PD* is identical to the Chao1-*PD* estimator proposed recently by Chao et al. (2015*a*), but the estimator of the shared Faith's *PD* (Eq. 10d) is new.
3. For functional trait diversity, the estimator of *FAD* in a single assemblage (Eq. 12d) and shared *FAD* between two assemblages (Eq. 14b) are both new; see Appendix S3: Tables S1–S3 for a summary.

The R code "Good-Turing" for computing all the estimators discussed in this paper is available in Github (https://github.com/AnneChao) along with a description of the running procedures. As an alternative, readers without a background in R, can utilize the online software "GoodTuring", made available from https://chao.shinyapps.io/GoodTuring/ to facilitate all computations.

We have proved that each of the derived estimators is theoretically a lower bound of the corresponding diversity. Good-Turing's perspectives reveal the sufficient conditions under which the resulting estimator is nearly unbiased. For example, a simple sufficient condition for the Chao1 species richness estimator being nearly unbiased is that *rare* species (specifically, undetected species and singletons in sample) have approximately homogenous abundances. Similar conditions for other estimators are also clearly specified in each subsection. See the next paragraph for more relaxed conditions.

Although, in our derivations, we follow the Good-Turing original model by assuming that the detection probability of each species is simply its relative abundance, all our derivations can be directly extended to a general model as discussed in Chao and Chiu (2016). The general model assumes that detection probability is proportional to the product of abundance and individual detectability, which may vary among species. Based on samples of individuals, individual detectability for mobile organisms is determined by many possible factors such as individual movement patterns, color, size, habitat, life cycle and vocalizations; for assemblages of sessile organisms, such as trees, that are surveyed from selected sampling units (e.g., transects, plots or quadrats), individual

detectability may depend on the area and topography of selected sampling units, species spatial/temporal aggregation or clustering, life history stage, as well as other factors. Consequently, our estimators are actually valid in more relaxed conditions. For example, the Chao1 species richness estimator is nearly unbiased when *hard-to-detect* species (specifically, undetected species and species detected by one individual) have approximately homogenous detection probabilities. Similar relaxed conditions can be formulated for other estimators. Under the special case that all individuals have the same detectability, the detection probability of each species reduces to its relative abundance. In other applications in which species detection probability may vary with time, our method can also be applied when the probability of detecting a species is modeled as the average detection probability over the sampling-time interval.

When rare or hard-to-detect species are highly heterogeneous in detection probabilities, such as in microbial assemblages or DNA sequencing data, all estimators derived in this paper provide non-parametric lower bounds which are valid for species, phylogenetic and functional diversities. In such assemblages, sample data do not provide sufficient information to accurately estimate asymptotic diversities due to heterogeneity of rare species; no statistical methods can produce reliable estimates unless strong assumptions are made. In such cases, from our perspective, an accurate lower bound is more practically useful than an imprecise point estimate. In Appendix S4, some representative simulation results are reported to validate the estimators derived from Good-Turing theory and also to demonstrate that our estimators provide useful and informative lower bounds when accurate point estimators are not attainable.

For cases that fail to meet or may not be assured to meet the criteria for nearly-unbiased point estimation, how can we make fair comparison of diversities across studies? We suggest using a non-asymptotic approach via sample-size- and coverage-based rarefaction and extrapolation on the basis of standardized sample size or sample completeness (as measured by sample coverage). This non-asymptotic approach facilitates fair comparison of diversities for equally-large or equally-complete samples across multiple assemblages. For species richness, sample-size-based rarefaction and extrapolation methods were developed by Colwell et al. (2012), and corresponding coverage-based methods were proposed by Chao and Jost (2012). For Faith's *PD*, a similar non-asymptotic method was recently presented by Chao et al. (2015*a*) and Hsieh and Chao (2017). Rarefaction and extrapolation methods for functional diversity are still under development by the authors.

For the analysis of multiple-assemblage phylogenetic structures, Webb et al. (2002) and Webb et al. (2008) proposed using a null model randomization test based on *PD* (or other metrics) to assess whether a specific assemblage has a higher *PD* (more even or uniform) or a lower *PD* (phylogenetic aggregation or clustering) than expected from an assembly randomly sampled from the *observed* species pool. That is, the observed species pool is the composite list of species from all assemblages; *PD* is calculated from the phylogenetic tree spanned by these detected species. In Webb et al.'s approach and in most null model or randomization tests (Gotelli et al. 2010 is an exception), it is assumed that (1) sampling is complete for all assemblages, and (2) the aggregation of observed species for all assemblages constitutes the "complete" species pool. However, for the Brazilian rain forest data analyzed in our analysis, if we treat each fragment as an assemblage, then neither of these assumptions is satisfied. Within each fragment, some species were undetected by the transect data, and some species remained undetected even in the pooled transect data for the entire study area represented by the 11 fragments. In the following, we use the Brazilian data to illustrate our suggested modifications.

1. Based on the pooled transect data from all 11 fragments, the Chao1 richness estimator yields the estimated number of species for the complete pool, which comprises all species recorded in the fragment transects plus undetected species in the study area represented by these fragments. Following the approach of Chao et al. (2015*b*), we can construct a complete species-rank abundance distribution (RAD) by separately adjusting the sample relative abundances for the set of species detected in the pooled, observed data and estimating the relative abundances for the set of species undetected in the pooled data, but inferred to be present in the area represented by all fragments. The combined RAD then fully characterizes the taxonomic assemblage structure. Thus, not only species richness but also species relative abundances can be estimated for all species in the estimated complete pool. For this analysis, it is not necessary to know the identities of the undetected species, but we do need to estimate their number and relative abundances in the complete pool.
2. Next, a random assemblage of the same size as each specific assemblage is sampled from the complete pool. If the sampled assemblage includes species that belong to the group of undetected species in the pooled data, then we treat those species as unresolved in their phylogenetic placement and locate those undetected species on the observed tree in some random manner (e.g., Rangel et al. 2015). How to optimally locate undetected species is still under investigation. The Chao1-*PD* estimate (Eq. 8d) can thus be calculated not only for the data of the specific assemblage but also for each random assemblage sampled from the complete pool.
3. After many assemblages of the same size have been randomly selected from the complete species pool as in Step (2), the mean and standard error of the resulting *PD* values can be computed to obtain the standardized effect size (Gotelli and McCabe 2002), adjusted for the under-sampling biases for both the specific assemblage and the pooled assemblage.

A similar procedure can be applied to *FAD* and other metrics, although this procedure should first be benchmarked with simulated and empirical data sets to assess its performance and to see if this procedure is an effective remedy for the under-sampling problem that is pervasive in the analysis of community structure based on standardized biodiversity sampling of multiple samples.

This paper is restricted to the estimation of diversity in one assemblage and the estimation of shared diversity between two assemblages. When there are more than two assemblages, our approach can be further extended to such cases. All the derivations are nearly parallel. For example, we can directly obtain an estimator of the species shared by multiple assemblages based on sampling data from each assemblage, and the resulting estimator is identical to the one proposed by Pan et al. (2009).

Our derivation in this paper is limited to individual-based abundance data for individuals sampled randomly from assemblages. In many ecological field studies, the sampling unit is not an individual, but a trap, net, quadrat, plot, or timed survey. For such studies, the sampling units, not the individuals, are sampled randomly and independently. In these cases, estimation is usually based on a set of sampling units in which only the incidence (detection or non-detection) of each species is recorded. This type of data is referred to as (multiple) incidence data. The sampling model and estimation were developed in Colwell et al. (2012) and Chao et al. (2014*b*). Chao and Colwell (2017) recently extended the Good-Turing frequency formula and its generalizations to incidence data. All estimators can be computed from the online software "GoodTuring". We expect that Good-Turing's theory will find wide applications in biodiversity studies; see Chao et al. (2017) for a recent application.

Acknowledgments

The authors thank a Subject Matter Editor (Tom Miller), Joaquín Hortal, and an anonymous reviewer for very thoughtful and helpful comments and suggestions. This work was supported by the Taiwan Ministry of Science and Technology under Contracts 104-2628-M007-003 and 105-2628-M007-001 (for AC) and 104-2118-M-002-008-MY3 (for CHC). RKC and RLC were supported by CAPES Ciência sem Fronteiras (Brazil). LFSM was supported by CAPES/PNPD. NJG was supported by U. S. NSF DEB 1257625, NSF DEB 1144055, and NSF DEB 1136644. The fieldwork survey was supported by Reserva Natural Vale, Fibria Celulose S.A., Marcos Daniel Institute, Pro-Tapir project and Reserva Biológica de Sooretama.

Literature Cited

Bryant, J. A., C. Lamanna, H. Morlon, A. J. Kerkhoff, B. J. Enquist, and J. L. Green. 2008. Microbes on mountainsides: contrasting elevational patterns of bacterial and plant diversity. Proceedings of the National Academy of Sciences of the United States of America 105:11505–11511.

Cadotte, M. W., J. Cavender-Bares, D. Tilman, and T. H. Oakley. 2009. Using phylogenetic, functional and trait diversity to understand patterns of plant community productivity. PLoS ONE 4:e5695.

Cardoso, P., F. Rigal, P. A. Borges, and J. C. Carvalho. 2014. A new frontier in biodiversity inventory: a proposal for estimators of phylogenetic and functional diversity. Methods in Ecology and Evolution 5:452–461.

Cavender-Bares, J., D. D. Ackerly, and K. H. Kozak. 2012. Integrating ecology and phylogenetics: the footprint of history in modern-day communities. Ecology 93:S1–S3.

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. 2005. Species estimation and applications. Pages 7907–7916 *in* N. Balakrishnan, C. Read, and B. Vidakovic, editors. Encyclopedia of statistical sciences. Wiley, New York.

Chao, A., and C.-H. Chiu. 2012. Estimation of species richness and shared species richness. Pages 76–111 *in* N. Balakrishnan, editor. Methods and applications of statistics in the atmospheric and earth sciences. Wiley, New York, USA.

Chao, A., and C.-H. Chiu. 2016. Species richness: estimation and comparison. Wiley StatsRef: Statistics Reference Online. 1–26.

Chao, A., and R. K. Colwell. 2017. Thirty years of progeny from Chao's inequality: estimating and comparing richness with incidence data and incomplete sampling (invited article). Statistics and Operation Research Transactions 41:3–54.

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

Chao, A., C.-H. Chiu, and L. Jost. 2014*a*. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. Annual Review of Ecology, Evolution, and Systematics 45:297–324.

Chao, A., N. J. Gotelli, T. C. Hsieh, E. L. Sander, K. Ma, R. K. Colwell, and A. M. Ellison. 2014*b*. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. Ecological Monographs 84:45–67.

Chao, A., C.-H. Chiu, T. C. Hsieh, T. Davis, D. A. Nipperess, and D. P. Faith. 2015*a*. Rarefaction and extrapolation of phylogenetic diversity. Methods in Ecology and Evolution 6:380–388.

Chao, A., T. C. Hsieh, R. L. Chazdon, R. K. Colwell, and N. J. Gotelli. 2015*b*. Unveiling the species-rank abundance distribution by generalizing the Good-Turing sample coverage theory. Ecology 96:1189–1201.

Chao, A., R. K. Colwell, C.-H. Chiu, and D. Townsend. 2017. Seen once or more than once: using Good-Turing theory to estimate species richness using only unique observations and a species list. Methods in Ecology and Evolution DOI: 10.1111/2041-210X.12768.

Chazdon, R. L., R. K. Colwell, J. S. Denslow, and M. R. Guariguata. 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of Northeastern Costa Rica. Pages 285–309 *in* F. Dallmeier, and J. A. Comiskey, editors. Forest biodiversity research, monitoring and modeling: conceptual background and old world case studies. Parthenon Publishing, Paris.

Chiu, C.-H., L. Jost, and A. Chao. 2014*a*. Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. Ecological Monographs 84:21–44.

Chiu, C.-H., Y. T. Wang, B. A. Walther, and A. Chao. 2014*b*. An improved nonparametric lower bound of species richness via a modified Good-Turing frequency formula. Biometrics 70:671–682.

Colwell, R. K., and J. A. Coddington. 1994. Estimating terrestrial biodiversity through extrapolation. Philosophical

Transactions of the Royal Society of London B - Biological Sciences 345:101–118.
Colwell, R. K., A. Chao, N. J. Gotelli, S.-Y. Lin, C. X. Mao, R. L. Chazdon, and J. T. Longino. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. Journal of Plant Ecology 5:3–21.
Diaz, S., and M. Cabido. 2001. Vive la différence: plant functional diversity matters to ecosystem processes. Trends in Ecology and Evolution 16:646–655.
Faith, D. P. 1992. Conservation evaluation and phylogenetic diversity. Biological Conservation 61:1–10.
Ferrier, S., G. Manion, J. Elith, and K. Richardson. 2007. Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distributions 13:252–264.
Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. Biometrika 40:237–264.
Good, I. J. 1983. Good thinking: the foundations of probability and its applications. University of Minnesota Press, Minneapolis, USA.
Good, I. J. 2000. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. Journal of Statistical Computation and Simulation 66:101–111.
Good, I. J., and G. Toulmin. 1956. The number of new species and the increase of population coverage when a sample is increased. Biometrika 43:45–63.
Gotelli, N. J., and A. Chao. 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. Pages 195–211 *in* S. A. Levin, editor. Encyclopedia of biodiversity. Second edition. Volume 5. Academic Press, Waltham, MA, USA.
Gotelli, N. J., and R. K. Colwell. 2011. Estimating species richness. Pages 39–54 *in* A. Magurran, and B. McGill, editors. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford.
Gotelli, N. J., and D. J. McCabe. 2002. Species co-occurrence: a meta-analysis of J.M. Diamond's assembly rules model. Ecology 83:2091–2096.
Gotelli, N. J., R. M. Dorazio, A. M. Ellison, and G. D. Grossman. 2010. Detecting temporal trends in species assemblages with bootstrapping procedures and hierarchical models. Philosophical Transactions of the Royal Society B 365:3621–3631.
Hortal, J., P. A. V. Borges, and C. Gaspar. 2006. Evaluating the performance of species richness estimators: sensitivity to sample grain size. Journal of Animal Ecology 75:274–287.
Hsieh, T. C., and A. Chao. 2017. Rarefaction and extrapolation: making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages. Systematic Biology 66:100–111.
Jesus, R. M., and S. G. Rolim. 2005. Fitossociologia da floresta atlântica de tabuleiro em Linhares (ES). Boletim Técnico SIF 19:1–149.
Jost, L., A. Chao, and R. Chazdon. 2011. Compositional similarity and beta diversity. Pages 66–84 *in* A. Magurran, and B. McGill, editors. Biological diversity: frontiers in measurement and assessment. Oxford University Press, Oxford, UK.
Lozupone, C., and R. Knight. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71:8228–8235.
Magnago, L. F. S., D. P. Edwards, F. A. Edwards, A. Magrach, S. V. Martins, and W. F. Laurance. 2014. Functional attributes change but functional richness is unchanged after fragmentation of Brazilian Atlantic forests. Journal of Ecology 102:475–485.
Magurran, A. E. 2004. Measuring biological diversity. Blackwell, Oxford, UK.
Matos, F. A. R., L. F. S. Magnago, M. Gastauer, J. M. B. Carreiras, M. Simonelli, J. A. A. Meira-Neto, and D. P. Edwards. 2017. Effects of landscape configuration and composition on phylogenetic diversity of trees in a highly fragmented tropical forest. Journal of Ecology 105:265–276.
McGrayne, S. B. 2011. The theory that would not die: how Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy. Yale University Press, New Haven, USA.
Pan, H. Y., A. Chao, and W. Foissner. 2009. A non-parametric lower bound for the number of species shared by multiple communities. Journal of Agricultural, Biological and Environmental Statistics 14:452–468.
Paula, A., and J. J. Soares. 2011. Estrutura horizontal de um trecho de Floresta Ombrófila Densa das Terras Baixas na Reserva Biológica de Sooretama, Linhares, ES. Revista Floresta 41:321–334.
Rangel, T. F., R. K. Colwell, G. R. Graves, K. Fučíková, C. Rahbek, and J. A. F. Diniz-Filho. 2015. Phylogenetic uncertainty revisited: Implications for ecological analyses. Evolution 69:1301–1312.
Robbins, H. E. 1968. Estimating the total probability of the unobserved outcomes of an experiment. The Annals of Mathematical Statistics 39:256–257.
Rolim, S. G., and H. E. M. Nascimento. 1997. Análise da riqueza, diversidade e relação espécie-abundância de uma comunidade arbórea tropical em diferentes intensidades amostrais. Scientia Forestalis 52:7–16.
Schmera, D., T. Erös, and J. Podani. 2009. A measure for assessing functional diversity in ecological communities. Aquatic Ecology 43:157–167.
Swenson, N. G., D. L. Erickson, X. Mi, N. A. Bourg, J. Forero-Montaña, X. Ge, R. Howe, J. K. Lake, X. Liu, and K. Ma. 2012. Phylogenetic and functional alpha and beta diversity in temperate and tropical tree communities. Ecology 93: S112–S125.
Tilman, D., J. Knops, D. Wedin, P. Reich, M. Ritchie, and E. Siemann. 1997. The influence of functional diversity and composition on ecosystem processes. Science 277: 1300–1302.
Walker, B., A. Kinzig, and J. Langridge. 1999. Plant attribute diversity, resilience, and ecosystem function: the nature and significance of dominant and minor species. Ecosystems 2:95–113.
Webb, C. O., and M. J. Donoghue. 2005. Phylomatic: tree assembly for applied phylogenetics. Molecular Ecology Notes 5:181–183.
Webb, C. O., D. D. Ackerly, M. A. McPeek, and M. J. Donoghue. 2002. Phylogenies and community ecology. Annual Review of Ecology and Systematics 33:475–505.
Webb, C. O., D. D. Ackerly, and S. W. Kembel. 2008. Phylocom: software for the analysis of phylogenetic community structure and trait evolution. Bioinformatics 24:2098–2100.

Supporting Information

Additional supporting information may be found in the online version of this article at http://onlinelibrary.wiley.com/doi/10.1002/ecy.2000/suppinfo