Methods in Ecology and Evolution

British Ecological Society

Methods in Ecology and Evolution 2014

doi: 10.1111/2041-210X.12247

SPECIAL FEATURE: NEW OPPORTUNITIES AT THE INTERFACE BETWEEN ECOLOGY AND STATISTICS

Rarefaction and extrapolation of phylogenetic diversity

Anne Chao[1]*, Chun-Huo Chiu[1], T. C. Hsieh[1], Thomas Davis[2], David A. Nipperess[2] and Daniel P. Faith[3]

[1]*Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan;* [2]*Department of Biological Sciences, Macquarie University, Sydney, NSW 2109, Australia; and* [3]*The Australian Museum, Sydney, NSW 2010, Australia*

Summary

1. Traditional species diversity measures do not make distinctions among species. Faith's phylogenetic diversity (*PD*), which is defined as the sum of the branch lengths of a phylogenetic tree connecting all species, takes into account phylogenetic differences among species and has found many applications in various research fields. In this paper, we extend Faith's *PD* to represent the total length of a phylogenetic tree from any fixed point on its main trunk.

2. Like species richness, Faith's *PD* tends to be an increasing function of sampling effort and thus tends to increase with sample completeness. We develop in this paper the '*PD* accumulation curve' (an extension of the species accumulation curve) to depict how *PD* increases with sampling size and sample completeness.

3. To make fair comparisons of Faith's *PD* among several assemblages based on sampling data from each assemblage, we derive both theoretical formulae and analytic estimators for seamless rarefaction (interpolation) and extrapolation (prediction). We develop a lower bound of the undetected *PD* for an incomplete sample to guide the extrapolation; the *PD* estimator for an extrapolated sample is generally reliable up to twice the size of the empirical sample.

4. We propose an integrated curve that smoothly links rarefaction and extrapolation to standardize samples on the basis of sample size or sample completeness. A bootstrap method is used to obtain the unconditional variances of *PD* estimators and to construct the confidence interval of the expected *PD* for a fixed sample size or fixed degree of sample completeness. This facilitates comparison of multiple assemblages of both rarefied and extrapolated samples.

5. We illustrate our formulae and estimators using empirical data sets from Australian birds in two sites. We discuss the extension of our approach to the case of multiple incidence data and to incorporate species abundances.

Key-words: diversity, extrapolation, phylogenetic diversity, rarefaction, sample coverage, species richness, undetected phylogenetic diversity

Introduction

Ecologists have developed various measures and approaches to quantify and compare biological diversities of assemblages; see Magurran & McGill (2011) for overviews. In traditional species diversity measures, all species are considered to be equally distinct from each other; only species richness and abundances are involved. Pielou (1975, p. 17) was the first to notice that traditional species diversity could be broadened to include phylogenetic, functional or other differences between species.

We here focus on phylogenetic differences among species. Such differences can be based directly on their evolutionary histories, either in the form of taxonomic classification or well-supported phylogenetic trees. There is a rapidly growing literature addressing phylogenetic diversity metrics and related similarity (or differentiation) measures; see a special issue in *Ecology* (Cavender-Bares, Ackerly & Kozak 2012) and papers in that issue.

The most widely used phylogenetic metric is Faith's phylogenetic diversity (*PD*) (Faith 1992a, b) which is defined as the sum of the branch lengths of a phylogenetic tree connecting all species in the target assemblage. Throughout the paper, Faith's *PD* is extended to refer to the total length of a phylogenetic tree from any point fixed in advance on its trunk, independent of the sampling results. If this reference point is chosen to be the root of the tree connecting all species, then it reduces to Faith's 1992 definition. As shown in Chao, Chiu & Jost (2010), *PD* can be regarded as a phylogenetic generalization of species richness. For example, when the branch lengths are proportional to divergence time so that all branch tips are the same distance from the root, *PD* divided by tree depth is referred to as 'lineage richness' in Chao, Chiu & Jost (2014).

Faith (1992) sees the link between *PD* and species in another way. *PD*, in counting-up branch lengths, is a proxy for 'feature

*Correspondence author. E-mail: chao@stat.nthu.edu.tw

diversity' (which is called 'attribute diversity' in Chao, Chiu & Jost 2014). Gains or losses in *PD* all can be interpreted as changes in the count of relative numbers of features under a model in which new features are generated in proportion to branch length. This analogy to counting-up species means that most ecological indices defined at the species level can be converted to *PD* equivalents (by counting features rather than species; Faith 2013).

For the past two decades, *PD* has found wide applications in various fields. The *PD* measure was developed for biodiversity conservation applications, where the goal is preservation of evolutionary history and feature diversity. Conservation priorities are calculated among threatened species on a phylogenetic tree, or among areas, recognizing that some places have high *PD* endemism or can complement an existing reserve system. In these applications, the major uncertainties involve estimation of phylogenetic topology, branch lengths and species distributions. Faith (1992a) also highlighted a fundamental source of uncertainty linked to *PD*'s assumption that shared ancestry explains shared features: 'PD values . . .may not be representative of the range of features among the taxa that are derived convergently'.

The concern in these applications is not so much estimating the true *PD* in a place or a region, but rather possible gains and losses of *PD* under different conservation scenarios. Several new directions in this rapidly expanding area of study now highlight the need for robust estimates of total *PD* of a community (or an assemblage), in contrast to the *PD* of a sample. Within community ecology, many studies now consider hypotheses about whether greater *PD* within the community or ecosystem corresponds to greater functionality, stability or resilience (e.g. Cadotte *et al*. 2009). Within microbial ecology, the range of hypotheses about total *PD* (typically for a molecular phylogeny of microbial variants) is even more varied and exciting. For example, Biedermann *et al*. (2013) examined the *PD* of the intestinal microbiota communities and found that people who had quit smoking had higher intestinal *PD* values. Schlaeppi *et al*. (2014) found that the *PD* of soil microbiota communities was greater for the roots of plants grown under natural conditions, compared with greenhouse-grown plants. These comparative *PD* studies naturally have highlighted the need to take sampling variation among sites/samples into account. In comparing the bacterial community-level *PD* across all 88 different soil types, Lauber *et al*. (2009) standardized using a randomly selected subset of 1200 molecular sequences. Similarly, Kembel *et al*. (2012) compared the microbial *PD* of different environments at a healthcare facility using *PD* based on samples 'rarefied' to 700 sequences per sample.

Like species richness, the empirical *PD* based on sampling data is highly dependent on sample size and sample completeness. When most species in an assemblage are rare, biodiversity samples are usually incomplete, so there are undetected species and thus undetected *PD* in samples. As a consequence, the observed *PD* is an underestimate of the true *PD*. The true *PD* refers to the *PD* of all species present in the entire assemblage, that is, it is the sum of the observed *PD* plus the undetected *PD*; note the true *PD* varies with the pre-selected reference point on the trunk.

The sampling-dependent problem has been extensively discussed for species richness. The traditional approach to control for this dependence in comparing the species richnesses of different communities is to use rarefaction to down-sample the larger samples until they are the same size as the smallest sample (e.g. see Gotelli & Colwell 2001). Ecologists then compare the richnesses of these equally large samples. However, this implies that some data in larger samples are thrown away. To avoid discarding data, Colwell *et al*. (2012) proposed using a sample-size-based rarefaction and extrapolation sampling curve that can be rarefied to smaller sample sizes or extrapolated to a larger sample size, guided by an estimate of asymptotic richness. Chao & Jost (2012) showed that rarefaction or extrapolation to a given degree of sample completeness (as measured by sample coverage; see later text) was better able to judge the magnitude of the differences in richness among communities, and ranked communities more efficiently, compared to traditional rarefaction and extrapolation to equal sample sizes. They developed coverage-based rarefaction and extrapolation methodology to implement this approach. The sample-size- and coverage-based integration of rarefaction and extrapolation together represent a unified framework for estimating species richness and for making statistical inferences based on these estimates.

Compared to species richness estimation and comparison, statistical estimation and standardization of phylogenetic diversity rarely have been explored. Nipperess & Matsen (2013) were the first to derive the exact analytic formula for the mean and (conditional) variance of *PD* under sample-size-based rarefaction. Cardoso *et al*. (2014) adapted some species richness estimators of taxon diversity to estimate *PD* of a community. However, there have been no extrapolation formulae for enlarged samples of finite sizes, nor any standardization by sample completeness.

In this paper, we first develop a '*PD* accumulation curve' to depict how *PD* increases with sampling size and sample completeness given any fixed point on its trunk. The proposed curve is an extension of the species accumulation curve to a phylogenetic version and can be used to construct rarefaction and extrapolation curves. We then generalize the sample-size-based approach of Colwell *et al*. (2012) and the coverage-based approach of Chao & Jost (2012) to phylogenetic versions, guided by an estimator of the undetected *PD*. Specifically, based on sampling data, we derive both theoretical formulae and analytic estimators for seamless rarefaction and extrapolation with *PD*. Our approach extends Nipperess & Matsen (2013) to include extrapolation for both sample-size- and coverage-based standardization.

To make fair comparison of *PD* among several assemblages based on sampling data from each assemblage, we propose an integrated curve that smoothly links rarefaction and prediction to standardize samples on the basis of sample size or sample completeness. A bootstrap method is used for obtaining the unconditional variance of *PD* estimators and constructing the confidence intervals of the expected *PD* for a fixed sample

size or fixed degree of sample completeness. This facilitates the comparison of multiple assemblages of both rarefied and extrapolated samples. We illustrate our formulae and estimators using empirical data sets from Australian birds in two sites.

Methods

PD ACCUMULATION CURVE

Consider an assemblage consisting of S species indexed by 1, 2, …, S. Let p_i denote the true relative abundance of species i, so that $\sum_{i=1}^{S} p_i = 1$. Assume that a rooted ultrametric or non-ultrametric phylogenetic tree of the S species (as tip nodes) can be constructed. As described in the *Introduction*, Faith's *PD* in this paper is extended to represent the total branch length of a phylogenetic tree from a predetermined reference point T_r on the main trunk. Then, the *PD* values from different samples (including samples with only one species) are all comparable. This extended *PD* is identical to the zero-order phylogenetic diversity $PD(T_r)$ introduced in Chao, Chiu & Jost (2010) for a reference point T_r on the trunk. Possible choices of the reference point include the root of the entire assemblage or the point of divergence between the focal organisms (e.g. birds) and their nearest outgroup; see Chiu, Jost & Chao (2014) for other possible choices of the reference points. As will be addressed in later text, all *PD* statistics for two different reference points on the trunk can be linked by a simple location transformation. For notational simplicity, we suppress the use of the reference point T_r in the following derivations.

Let $\boldsymbol{B} = \boldsymbol{B}_{T_r}$ denote the set of all branches given any fixed point T_r on its main trunk. Let L_i denote the length of branch i, and a_i denote the true total relative abundance descended from branch i, $i \in \boldsymbol{B}$. Thus, the set $(p_1, p_2, \ldots, p_S)$ for the S species (as tip nodes) is expanded to the set $\{a_i,\ i \in \boldsymbol{B}\}$ with $(p_1, p_2, \ldots, p_S)$ as the first S elements; see the left panels of Fig. 1 for simple examples and notation. We refer to $a_i,\ i \in \boldsymbol{B}$, as the *branch relative abundance* of branch i. The true *PD* is expressed as

$$PD = \sum_{i \in \boldsymbol{B}} L_i. \qquad \text{eqn 1}$$

When a sample of m individuals is taken from the assemblage, let $PD(m)$ denote the expected *PD* of this sample. The formula for $PD(m)$ is derived in Appendix S1

$$PD(m) = \sum_{i \in \boldsymbol{B}} L_i[1 - (1 - a_i)^m], \quad m = 1, 2, \ldots \qquad \text{eqn 2}$$

This shows that $PD(m)$ is a non-decreasing function of sample size. The plot of $PD(m)$ as a function of m is the *sample-size-based PD accumulation curve*. As sample size m tends to infinity, $PD(m)$ approaches the true *PD*. Thus, the true *PD* represents the 'asymptote' of the *PD* accumulation curve, that is the true $PD = PD(\infty)$. This is analogous to species richness in that the true species richness represents the asymptote of the species accumulation curve. When there are no internal nodes and all S branches are equally distinct with branch lengths of unity (i.e. branch lengths are normalized to unity), our *PD* accumulation curve reduces to the species accumulation curve.

Chao & Jost (2012) proposed standardizing samples by sample completeness, which is measured by *sample coverage* (or simply *coverage*), a concept originally developed by the founder of modern computer science, Alan Turing and I. J. Good (Good 1953, 2000). The sample coverage of a given sample is defined as the proportion of the total number of individuals in an assemblage that belong to the species represented in the sample. Sample coverage is also a function of sample size. Let $C(m)$ be the expected sample coverage for a sample of size m. The plot of $PD(m)$ as a function of $C(m)$ is the *coverage-based PD accumulation curve*. As $C(m)$ tends to unity (complete coverage), the curve also approaches the true *PD*.

A REFERENCE SAMPLE OF SIZE *N*

We assume an empirical sample of n individuals is taken *with replacement* from the assemblage, and a total of S_{obs} ($\le S$) species are observed. Following the terminology of Colwell *et al.* (2012), we call this sample the *reference sample*. Let X_i be the number of individuals of the ith species that are observed in the sample, $i = 1, 2, \ldots, S$; we refer to X_i as the *sample species frequency*. Let f_k be the number of species represented by exactly k individuals in the reference sample, $k = 0, 1, \ldots, n$; we refer to $\{f_k,\ k = 0,\ 1,\ \ldots,\ n\}$ as the sample *abundance frequency counts*. From these definitions, $n = \sum_{i=1}^{S} X_i = \sum_{k \ge 1} k f_k$, and $S_{obs} = \sum_{k \ge 1} f_k$. In particular, f_1 is the number of species represented by exactly one individual (*singletons*) in the sample, and f_2 is the number of species represented by exactly two individuals (*doubletons*). The unobservable frequency f_0 denotes the number of species that are present in the entire assemblage but are not detected in the reference sample.

Given the reference sample of size n and species true relative abundances $(p_1, p_2, \ldots, p_S)$ of the assemblage, the sample species frequencies $(X_1,\ X_2,\ \ldots,\ X_S)$ follow a multinomial probability distribution

$$P(X_1 = x_1, \ldots, X_S = x_S) = \frac{n!}{x_1! \ldots x_S!} p_1^{x_1} p_2^{x_2} \ldots p_S^{x_S}. \qquad \text{eqn 3}$$

Although the undetected species, that is $X_i = 0$, do not contribute to this distribution, eqn 3 provides a model to infer the number of undetected species; see below. Based on the observed species in the reference sample, we can draw the observed phylogenetic tree, which is a portion of the entire phylogenetic tree for all species; see the right panels of Fig. 1 for illustration. Let PD_{obs} denote the observed *PD* in the reference sample. We now expand the set of sample species frequencies $(X_1,\ X_2,\ \ldots, X_S)$ to a larger set $\{X_i,\ i \in \boldsymbol{B}\}$ by defining X_i, $i \in \boldsymbol{B}$, as the sum of the sample species frequencies for those species descended from branch i. We refer to $X_i,\ i \in \boldsymbol{B}$, as the *sample branch frequency (or abundance)* of branch i. Under a multinomial model, the sample branch frequency X_i of branch i follows a binomial distribution

$$P(X_i = x_i) = \binom{n}{x_i} a_i^{x_i} (1 - a_i)^{n - x_i}, \quad \text{for } i \in \boldsymbol{B}. \qquad \text{eqn 4}$$

Based on a reference sample of size n with the sample branch frequencies $\{X_i,\ i \in \boldsymbol{B}\}$, our purpose is to provide the formulae and estimators for $PD(m)$, $m = 1, 2, \ldots$. Rarefaction refers to the case $m < n$, whereas extrapolation refers to the case $m > n$; the two parts join at the reference sample size n. The integrated sample-size- or coverage-based rarefaction and extrapolation sampling curve represents the estimated *PD* accumulation curve based on the reference sample.

RAREFACTION

When $m \le n$, statistical estimation theory shows that the unique minimum variance unbiased estimator exists for $PD(m)$, the expected *PD* for a sample of size m (given in eqn 2). The estimator is (Appendix S1 for proof)

$$\widehat{PD}(m) = PD_{obs} - \sum_{\substack{i \in \boldsymbol{B} \\ 1 \le X_i \le n - m}} L_i \frac{\binom{n - X_i}{m}}{\binom{n}{m}}, \quad m \le n. \qquad \text{eqn 5}$$

For any fixed $m < n$, the right-most term in the above equation is non-zero if there are species with sample frequencies $1 \le X_i \le n - m$.

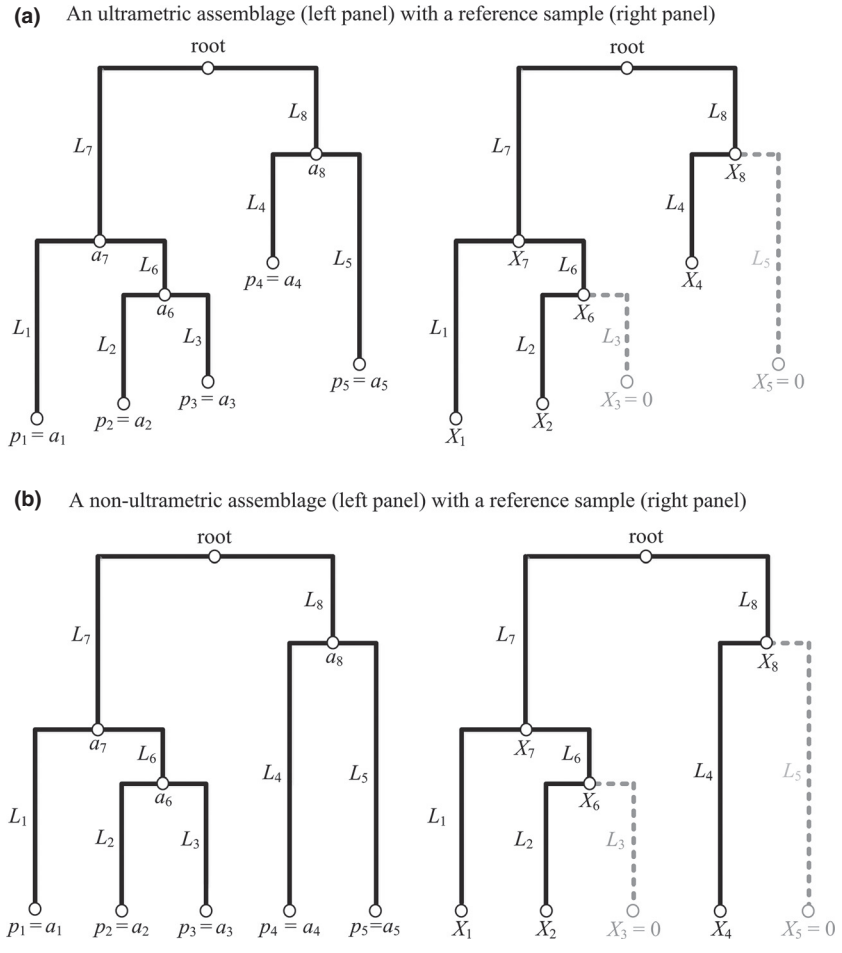

Fig. 1. (a) A hypothetical ultrametric assemblage with a reference sample. (b) A hypothetical non-ultrametric assemblage with a reference sample. The ancestor of the entire assemblage is the 'root' at the top, with time (or base change) progressing towards the branch tips at the bottom. Here the root of the entire assemblage is selected as the reference point for illustration. (Any other point on the trunk could also be selected as a reference point.) Each internal node (branching point) represents a speciation or divergence event, and the 5 branch tips illustrate the 5 extant species indexed by 1, 2, …, 5 with true relative abundances $(p_1, p_2, \ldots, p_5)$. The branch set $\boldsymbol{B}$ includes 8 branches (indexed from 1 to 8) with branch lengths $(L_1, L_2, \ldots, L_8)$, and the corresponding branch relative abundances $(a_1, a_2, \ldots, a_8) = (p_1, p_2, p_3, p_4, p_5, p_2 + p_3, p_1 + p_2 + p_3, p_4 + p_5)$ with $(p_1, p_2, \ldots, p_5)$ as the first 5 elements. In the reference sample, species 3 and species 5 are not detected, so only a portion of the tree (solid branches in the right panels) is observed and the dotted branches (L_3 and L_5) are not detected in the reference sample. The sample species frequencies for the five species (tip nodes) are $(x_1, x_2, 0, x_4, 0)$. The sample branch frequencies for all 8 branches take the specific value of $(x_1, x_2, 0, x_4, 0, x_2, x_1 + x_2, x_4)$, that is $x_6 = x_2$, $x_7 = x_1 + x_2$, $x_8 = x_4$.

When $m = n$, the right-most term is an empty sum and thus the above formula reduces to the observed *PD* in the reference sample, that is $\widehat{PD}(n) = PD_{obs}$.

When there are no internal nodes and all S branches are equally distinct with a normalized branch length of unity, eqn 5 reduces to the traditional rarefaction formula for species richness. A bootstrap method is used to obtain the unconditional variance estimation of $\widehat{PD}(m)$ and the confidence interval of $PD(m)$; see Appendix S2 for a description of the bootstrap method.

When sampling is conducted by selecting individuals without replacement so that no individuals can be repeatedly sampled, a slight modification is needed for the corresponding statistical model and *PD* accumulation curve, but the rarefaction estimator has the same form as that in eqn 5; see Appendix S2 for details.

EXTRAPOLATION

As shown in Colwell *et al.* (2012) and Chao & Jost (2012), the extrapolation formula for species richness requires a predictor for the undetected number of species in the reference sample, that is, f_0, the number of species with species frequency 0. As $S = S_{obs} + f_0$, it is equivalent to requiring a species richness estimator. Colwell *et al.* (2012) and Chao & Jost (2012) adopted the use of the Chao1 estimator (Chao 1984), which is a theoretical lower bound of species richness under a commonly used multinomial model (eqn 3). Statistically, it is difficult to accurately estimate species richness if there are many almost undetectable species in a hyper-diverse assemblage. Based on a wide range of simulation scenarios, an accurate lower bound for species richness is preferable to an inaccurate point estimator (Chiu *et al.* 2014). The Chao (1984) estimator uses only the information on rare species (numbers of singletons and doubletons) to estimate the number of undetected species in samples

$$\hat{S}_{Chao1} = \begin{cases} S_{obs} + \dfrac{(n-1)}{n}\dfrac{f_1^2}{(2f_2)}, & \text{if } f_2 > 0\,; \\ S_{obs} + \dfrac{(n-1)}{n}\dfrac{f_1(f_1-1)}{2}, & \text{if } f_2 = 0\,. \end{cases} \qquad \text{eqn 6}$$

The formula for the case $f_2 = 0$ (e.g. Chao 2005, p. 7910; Chao & Jost 2012; their eqn 8) is a bias-corrected estimator under a homogeneous model; see Appendix S2 for more details. Cardoso *et al.* (2014) adapted the above species richness estimator to make inferences for *PD*. However, their adaptation is not theoretically proved for *PD* data. As they concluded, the development of estimators specifically for *PD* data is needed. Here for the first time, we derive a theoretical lower bound for the undetected *PD* under our phylogenetic framework. Define g_k as the sum of branch lengths for those branches with sample branch frequency k, that is

$$g_k = \sum_{i \in \boldsymbol{B}} L_i \, I(X_i = k), \;\; k = 0, \; 1, \ldots,$$

where $I(\cdot)$ is an indicator function that equals 1 when true and 0 otherwise. The undetected *PD* in the reference sample is g_0, which is the total length of undetected branches in the reference sample; g_0 is unknown but $\{g_1, g_2, \ldots\}$ can be computed from the reference sample and the tree spanned by the observed species. Here g_1

denotes the total branch length of those singletons in the set $\{X_i,\ i \in \boldsymbol{B}\}$, and g_2 denotes the total branch length of those doubletons in the set $\{X_i,\ i \in \boldsymbol{B}\}$.

The expected value of g_k can be expressed as

$$E(g_k) = E\sum_{i\in\boldsymbol{B}} L_i I(X_i = k) = \binom{n}{k} \sum_{i\in\boldsymbol{B}} L_i a_i^k (1-a_i)^{n-k}. \qquad \text{eqn 7}$$

In particular, we have $E(g_0) = E\sum_{i\in\boldsymbol{B}} L_i I(X_i = 0) = \sum_{i\in\boldsymbol{B}} L_i (1-a_i)^n$. Also, we have the following two expectation formulae:

$$E(g_1) = E\sum_{i\in\boldsymbol{B}} L_i I(X_i = 1) = n\sum_{i\in\boldsymbol{B}} L_i a_i (1-a_i)^{n-1}.$$

$$E(g_2) = E\sum_{i\in\boldsymbol{B}} L_i I(X_i = 2) = \frac{n(n-1)}{2}\sum_{i\in\boldsymbol{B}} L_i a_i^2 (1-a_i)^{n-2}.$$

The Cauchy-Schwarz inequality leads to

$$\left(\sum_{i\in\boldsymbol{B}} L_i (1-a_i)^n\right) \times \left(\sum_{i\in\boldsymbol{B}} L_i a_i^2 (1-a_i)^{n-2}\right) \geq \left(\sum_{i\in\boldsymbol{B}} L_i a_i (1-a_i)^{n-1}\right)^2,$$

with equality if $a_i,\ i \in \boldsymbol{B}$, is a constant. The first term in the left hand side of the above inequality is $E(g_0)$, the second term is $2E(g_2)/[n(n-1)]$, and the term on the right side is $[E(g_1)/n]^2$. Thus, the above inequality gives a lower bound for $E(g_0)$

$$E(g_0) \geq \frac{(n-1)}{n}\frac{[E(g_1)]^2}{2E(g_2)}. \qquad \text{eqn 8}$$

Replacing the expected value by the observed value, we have the following lower bound for the undetected *PD*:

$$\tilde{g}_0 = \frac{(n-1)}{n}\frac{g_1^2}{2g_2}, \quad \textit{if } g_2 > 0. \qquad \text{eqn 9}$$

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. In this case, we propose the following modified lower bound for the undetected *PD* based on simulation results in Appendix S3:

$$\hat{g}_0 = \begin{cases} \dfrac{(n-1)}{n}\dfrac{g_1^2}{2g_2}, & \textit{if } g_2 > \dfrac{g_1 f_2}{2f_1}; \\ \dfrac{(n-1)}{n}\dfrac{g_1(f_1-1)}{2(f_2+1)}, & \textit{if } g_2 \leq \dfrac{g_1 f_2}{2f_1}. \end{cases} \qquad \text{eqn 10}$$

Note that when $n > 2$, the estimator $\hat{g}_0$ is independent of the choice of the reference point on the trunk. This is because the sample branch frequency for the trunk segment is the sample size n; this frequency of n does not contribute to g_1 and g_2 if $n > 2$ and thus does not have any effect on the estimator $\hat{g}_0$. We also propose the following Chao1-*PD* lower bound (or estimator) for the true *PD* based on a reference sample of size n:

$$\widehat{PD}_{Chao1} = PD_{obs} + \hat{g}_0 = \begin{cases} PD_{obs} + \dfrac{(n-1)}{n}\dfrac{g_1^2}{2g_2}, & \textit{if } g_2 > \dfrac{g_1 f_2}{2f_1}; \\ PD_{obs} + \dfrac{(n-1)}{n}\dfrac{g_1(f_1-1)}{2(f_2+1)}, & \textit{if } g_2 \leq \dfrac{g_1 f_2}{2f_1}. \end{cases} \qquad \text{eqn 11}$$

This lower bound is a nearly unbiased estimator if (i) branch abundances are homogeneous, or (ii) sample size is sufficiently large. When branch abundances are heterogeneous and sample size is not sufficiently large, negative bias exists. Nevertheless, simulation results (Appendix S3) suggested the use of the Chao1-*PD* lower bound as an estimator of the true *PD* can be recommended. Therefore, we also refer to it as an estimator throughout the paper. A bootstrap method (described in Appendix S2) or a standard asymptotic approach can be used to approximate the variance of the proposed *PD* estimator and obtain the associated confidence intervals.

In a sample-size-based sampling curve, given the data for a reference sample of size n, the extrapolation problem is to predict the expected *PD* in an augmented sample of $n + m^*$ individuals from the assemblage ($m^* \geq 0$). Based on the theoretical formula given in eqn. 2 for a sample size of $n + m^*$, we obtain (see Appendix S4 for proof) the extrapolated *PD* lower bound (or estimator) at sample size $n + m^*$

$$\widehat{PD}(n+m^*) = PD_{obs} + \hat{g}_0\left[1 - \left(1 - \frac{g_1}{n\hat{g}_0 + g_1}\right)^{m^*}\right],\ m^* \geq 0, \qquad \text{eqn 12}$$

where $\hat{g}_0$ is an estimator for g_0 (the undetected *PD*) given in eqn 10. A bootstrap method (described in Appendix S2) is used to obtain the unconditional variance estimate of $\widehat{PD}(n+m^*)$ and the associated confidence interval. We summarize the theoretical formula for $PD(m)$, $m > 0$, and analytic estimators for rarefaction and extrapolation in Table 1. For the special case of $m^* = 0$, the extrapolated estimator reduces to the observed *PD* in the reference sample. Thus, rarefaction and extrapolation curves seamlessly join at the reference sample point. See Appendix S5 for the performance of our proposed rarefaction and extrapolation estimators based on simulations.

INTEGRATED CURVES

Our proposed sample-size-based sampling curve for *PD* includes the rarefaction part (which plots $\widehat{PD}(m)$ as a function of m, where $m \leq n$) and the extrapolation part (which plots $\widehat{PD}(n+m^*)$ as a function of $n + m^*$, $m^* \geq 0$) and yields a smooth sampling curve, the two parts of which join smoothly at the point of the reference sample (n, PD_{obs}). The confidence intervals based on the bootstrap method also join smoothly. See the *Application* section for examples.

The expected coverage $C(m)$ (defined in *PD accumulation curve*) for a sample size of m is an increasing function of m. The theoretical formula (Chao & Jost 2012) is given in Table 1. To construct a coverage-based sampling curve, we need an interpolated coverage estimator $\hat{C}(m)$ for any rarefied sample of size $m < n$ and an extrapolated coverage estimator $\hat{C}(n+m^*)$ for any augmented sample of size $n + m^*$. Chao & Jost (2012) derived such estimators; see Table 1 for formulae.

Our coverage-based sampling curve for *PD* includes rarefaction (which plots $\widehat{PD}(m)$ with respect to $\hat{C}(m)$) and extrapolation (which plots $\widehat{PD}(n+m^*)$ with respect to $\hat{C}(n+m^*)$) joining smoothly at the reference sample point ($\hat{C}(n)$, PD_{obs}). The confidence intervals based on the bootstrap method also join smoothly. Note that the pattern for each of the two types (sample-size- and coverage-based) of integrated curves is not affected by the choice of the reference point on the trunk. See *Application* for examples.

As proposed by Chao *et al.* (2014), the sample-size- and coverage-based sampling curves can be bridged by the relationship between sample coverage and sample size. Using the coverage estimators in Table 1 (the last row), we can construct a *sample completeness curve*, which plots the estimated sample coverage with respect to sample size. From the reference sample, this curve estimates sample completeness for smaller rarefied samples, as well as for larger extrapolated samples. See the next section for examples.

Table 1. The theoretical formulae and analytic estimators for rarefaction and extrapolation of *PD*, given a reference sample with the observed *PD*, PD_{obs} and the estimated coverage $\hat{C}(n) = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1+2f_2}\right]$ (if $f_2 > 0$). (See Appendix S2 for $\hat{C}(n)$ if $f_2 = 0$). The last row gives formulae for the expected sample completeness as a function of sample size, and the corresponding coverage estimators for rarefied samples and extrapolated samples

Theoretical formula (for $m > 0$)	Interpolation estimator (for a sample of size $m < n$)	Extrapolation estimator (for a sample of size $n + m^*$, $m^* \geq 0$)
Expected PD: $PD(m) = \sum_{i \in \boldsymbol{B}} L_i[1 - (1 - a_i)^m]$	$\widehat{PD}(m) = PD_{obs} - \sum_{\substack{i \in \boldsymbol{B} \\ 1 \leq X_i \leq n-m}} L_i \frac{\binom{n-X_i}{m}}{\binom{n}{m}}$ (unbiased estimator)	$\widehat{PD}(n+m^*) = PD_{obs} + \hat{g}_0\left[1 - \left(1 - \frac{g_1}{n\hat{g}_0 + g_1}\right)^{m^*}\right]$ (reliable if $m^* < n$)
Expected coverage: $C(m) = 1 - \sum_{i=1}^{S} p_i(1-p_i)^m$	$\hat{C}(m) = 1 - \sum_{\substack{i=1 \\ 1 \leq X_i \leq n-m}}^{S} \frac{X_i}{n} \frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}$ (unbiased estimator)	$\hat{C}(n+m^*) = 1 - \frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1+2f_2}\right]^{m^*+1}$ (reliable if $m^* < n$)

Application

We use bird data collected in November 2012 at Barrington Tops National Park, Australia, for illustration. The data were collected from 29 points, and at each point, abundances of the total number of birds observed over a 30-min period in a 50 m radius were recorded. There are 17 points along the Gloucester Tops Road in the southern part of the Barrington Tops National Park, and the sample species frequencies of these 17 points are pooled to form the reference sample for 'South-site'. There are 12 points along the Barrington Tops Forest Road in the northern part of the National Park, and the frequencies are pooled to form the reference sample for 'North-site'; see Fig. S6·1 (Appendix S6) for species sample frequencies. Vegetation at both sites ranged from wet Eucalypt forest to rain forest with an average canopy cover of 60% for South-site and 80% for North-site. The sampling points comprising South-site had an average elevation of 928 m while those of North-site had an average elevation of 1078 m.

A total of 41 species were observed. A phylogenetic tree of these species was constructed from a Maximum Clade Credibility tree of the Bayesian analysis of Jetz *et al.* (2012); see Appendix S6. The age of the root for 41 species is 82·9 million years (Myr). Without loss of generality, we selected the time depth at 82·9 Myr as our temporal perspective in our statistical analysis for illustration. Although the root of the observed species varies with sampling data, we can easily transform all our estimates to those for a new reference point on the trunk. For example, if we change to a new time depth at 82·9 + *A* Myr (and thus the true *PD* is increased by *A* Myr) to compare results across studies, then all our rarefaction/extrapolation estimates and the Chao1-*PD* estimate of the true *PD* (eqn 11) for this new perspective will be increased by the same magnitude of *A* Myr; see Appendices S1, S3 and S4 for details. Generally, as long as sample size is larger than 2, the pattern of our rarefaction/extrapolation curves is invariant to the choice of the reference time depth because the true PD and each of our statistics for two reference points are linked via a simple location transformation.

The species abundance frequency counts for the two sites are tabulated in Table S6·1 (Appendix S6) along with some statistics that will be explained below. We use the data from these two sites to illustrate the construction of two types of rarefaction and extrapolation curves (sample-size- and coverage-based) with *PD* (Figs 2 and 4), and the sample completeness curve (Fig. 3). The constructed sampling curves are then used to compare the *PD* between the two plots.

The reference sample (with a sample size of 307) in South-site includes 38 species with the observed *PD* (PD_{obs}) of 1416·7 Myr. The sample completeness is high, as reflected by our coverage estimator (Table 1) $\hat{C}(n)$ = 98·4%. The total

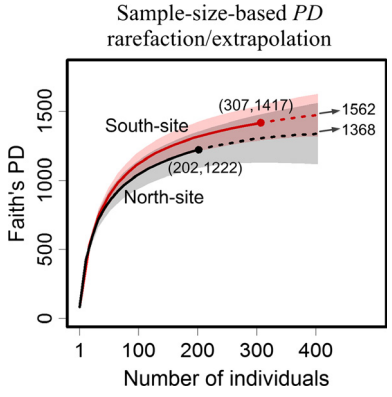

Fig. 2. Comparison of sample-size-based rarefaction (solid lines) and extrapolation (dotted curves) for *PD*, up to the base sample size of 400 individuals (i.e. double the smaller reference sample size) for Australian bird species data in South-site and North-site. The fixed time depth is 82·9 Myr (the age of the root of the observed tree). Reference samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. The numbers in parentheses are the sample size and the observed *PD* for each reference sample. The estimated asymptote of *PD* (eqn 11) for each curve is shown after an arrow sign.

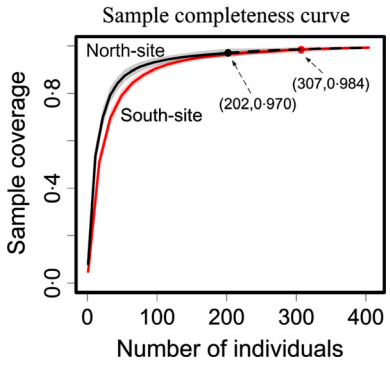

Fig. 3. Plot of sample coverage for rarefied samples (solid line) and extrapolated samples (dashed line) as a function of sample size for Australian bird species data in two sites. Reference samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. Each of the two curves was extrapolated up to the base sample size of 400. The numbers in parentheses are the sample size and the estimated sample coverage for each reference sample.

branch lengths for singletons in the sample branch frequencies are calculated as g_1 = 222·43 Myr, and for doubletons is g_2 = 170·36 Myr. These two branch lengths produce (by eqn 10) an estimate of the undetected *PD* as $\hat{g}_0$ = 145·22 Myr, leading to the Chao1-*PD* estimate of the true *PD* (by eqn 11), $\widehat{PD}_{Chao1} = PD_{obs} + \hat{g}_0$ = 1561·94 (with an estimated SE of 230·84 based on a bootstrap method of 200 replications).

The reference sample (with a sample size of 202) in North-site includes 27 species with PD_{obs} = 1215·98. The coverage of the sample is estimated to be $\hat{C}(n)$ = 97·0%. For this site, we have g_1 = 231·80, g_2 = 183·57, leading to an estimate of the undetected *PD* (by eqn 10), $\hat{g}_0$ = 146·35 and the Chao1-*PD* estimate of the true *PD* (by eqn 11) $\widehat{PD}_{Chao1} = PD_{obs} + \hat{g}_0$ = 1368·45 (with an estimated SE of 276·03 based on a bootstrap method of 200 replications). These estimates of undetected *PD*s are required in calculating our extrapolation formula.

Here we apply the procedures proposed in Chao *et al.* (2014) to compare standardized samples between the two sites as follows:

STEP (1): COMPARE SAMPLE-SIZE-BASED SAMPLING CURVES UP TO A BASE SAMPLE SIZE

We first compare the integrated sample-size-based rarefaction and extrapolation curves for *PD* along with 95% confidence intervals (based on a bootstrap method of 200 replications) up to a *base sample size* of 400. Here the base sample size is defined to be double the smaller reference sample size, as suggested by Chao *et al.* (2014). The estimated *PD* and confidence intervals then can be compared across sites for any sample size less than the base size. Across this range of abundance, Fig. 2 reveals that South-site has higher *PD* than that of North-site, but the two confidence intervals overlap. Generally, for any fixed sample size (or completeness) in the comparison range, if the 95% confidence intervals do not overlap, then significant differences at a level of 5% among the expected diversities (whether interpolated or extrapolated) are guaranteed. However, overlapped intervals do not guarantee non-significance (Colwell *et al.* 2012).

STEP (2): CONSTRUCT A SAMPLE COMPLETENESS CURVE TO LINK SAMPLE-SIZE- AND COVERAGE-BASED SAMPLING CURVES

As discussed earlier, the estimated coverages for the reference sample in South-site and North-site are, respectively, 98·4% and 97·0%. However, these two coverage values are based on different reference sample sizes. Figure 3 plots how the sample completeness varies with sample size along with 95% confidence intervals for each of the two sites, up to the base sample size of 400. The curve shows for any fixed size $\leq$ 400 that the sample completeness for North-site is estimated to be consistently higher than that in South-site. When sample size is larger than 200, the sample coverage estimates for the two sites differ very little. The sample completeness curve provides a bridge between sample-size- and coverage-based sampling curves.

STEP (3): COMPARE COVERAGE-BASED SAMPLING CURVES UP TO A BASE COVERAGE

From the sample completeness curve (Fig. 3), when sample size in North-site is doubled from 202 to 400 individuals, the sample coverage is increased from 97% to 99·20%. In South-site, when sample size is increased from 307 to the double size 600, the coverage is increased from 98·4% to 99·8%. In Fig. 4, we compare the corresponding coverage-based rarefaction and extrapolation curves of *PD* with 95% confidence intervals up to the coverage of 99·2%. This is our '*base coverage*' (the lower of the two coverages for the doubled reference sample sizes). Because the increase in coverage for the extrapolation is small, and the estimated *PD* hardly changes beyond the reference samples, the extrapolation parts in Fig. 4 are nearly invisible. An enlarged plot for coverage > 0·8 is thus shown in Fig. 4.

The enlarged plot reveals that the two confidence bands do not intersect for *PD* if coverage < 90% (except for the initial stages). This is one advantage of using coverage-based curves: data show that the *PD* in South-site is significantly higher than that in North-site for any standardized sample coverage less than 90%. When the coverage is higher than 90%, the *PD* ordering remains the same but confidence intervals intersect. This may be due to wider confidence intervals or convergent species composition when sample coverage is increased.

Discussion

In this paper, we have developed a novel statistical framework for the analysis of biodiversity data based on *PD*. We propose constructing two types (sample-size- and coverage-based) of integrated rarefaction and extrapolation curves as illustrated in Figs 2 and 4. These curves are then used to compare *PD* values among multiple assemblages for standardized sample size or sample completeness. The sample-size- and coverage-based curves are linked by a sample completeness curve (Fig. 3),

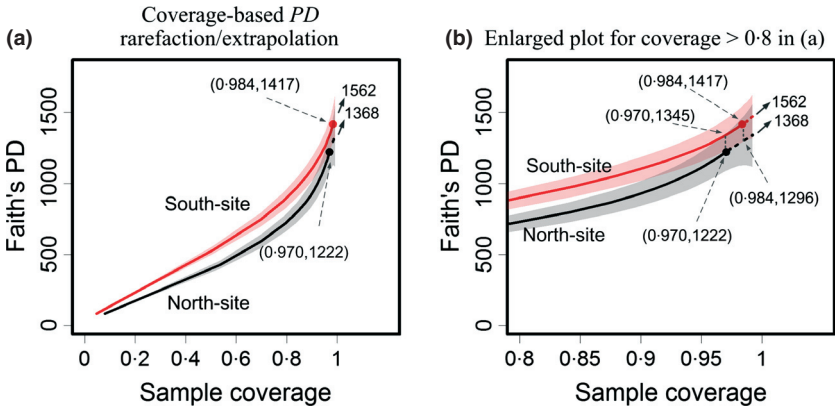

Fig. 4. (a) Comparison of the coverage-based rarefaction (solid lines) and extrapolation (dotted curves), up to the base coverage 99·2% (the lower coverage of the doubled reference sample sizes) for Australian bird species data in South-site and North-site. The fixed time depth is 82·9 Myr (the age of the root of the observed tree). Reference samples are denoted by solid dots. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. The numbers in parentheses are the sample coverage and the observed *PD* (for solid dots) or estimated *PD* (otherwise). The estimated asymptote of *PD* (eqn 11) for each curve is shown after an arrow sign. (b) Enlarged plot for coverage > 0·8 in (a).

which reveals the relationship between sample size and sample completeness. See Paulson *et al.* (2013) for a different standardization method.

In our example application, two bird communities (North-site and South-site) were compared for phylogenetic diversity. The data from these two communities differed in their sample sizes (number of individuals) and coverages. When standardized to the same sample coverage, South-site had consistently greater *PD* than North-site, this difference becoming significant when coverage was less than 90%. Due to lower average canopy cover, sampling points in the South-site generally had denser middle and lower story vegetation, providing habitat for small-to-medium forest birds such as the Superb Fairywren (*Malurus cyaneus*). The South-site also had a lower average elevation. Species richness of birds is generally observed to decline with increasing elevation, possibly due to a decline in available energy (Williams *et al.* 2010). Thus, the differences between the sites in *PD* may simply reflect differences in species richness, which itself may be influenced by canopy cover and elevation. With most of the phylogenetic tree being represented at both sites (Fig. S6·1 of Appendix S6), the increased *PD* of South-site seems to be driven mostly by a larger number of terminal taxa, rather than being a particularly phylogenetically distinctive assemblage.

Our experiences suggest for *PD* that the proposed estimators work well for rarefaction and short-range extrapolation in which the extrapolated sample size is up to twice the reference sample size. This finding is consistent with that for rarefaction and extrapolation for species richness (Chao & Jost 2012; Colwell *et al.* 2012). For rarefaction, our proposed estimator is unbiased. When the extrapolated sample size is more than double the reference sample size, the magnitude of the prediction bias generally increases with the prediction range, and the extrapolation is reliable up to no more than double the reference sample size. Beyond that, the predictor may be subject to some bias because our asymptotic estimator for the undetected *PD* (eqn 10) theoretically is a lower bound only.

Our presentation here is focused on the sampling curve when a sample of individuals is taken from an assemblage. Gotelli & Colwell (2001) distinguished two types of rarefaction curves: individual-based (the sampling unit is an individual) and sampling-unit-based (the sampling unit is a sample or quadrat and only species incidences are recorded). Our unified *PD* rarefaction/extrapolation approach for standardizing sample size and coverage can be extended to handle multiple incidence data. This extension is presented in Appendix S7, where we also propose a Chao2-*PD* lower bound of the true *PD* based on incidence data.

Chao, Chiu & Jost (2010) extended *PD* to incorporate species abundance based on a framework of Hill numbers (Hill 1973). They proposed a class of phylogenetic diversity with an order q. Faith's *PD* for a fixed reference point represents the phylogenetic diversity of order zero. Their measures were justified and extended by Faith (2013). Both the original Hill numbers and their phylogenetic generalizations facilitate diversity decomposition (Chiu, Jost & Chao 2014). Recently, Chao *et al.* (2014) have developed rarefaction/extrapolation with Hill numbers. We are currently working on the rarefaction/extrapolation for the class of phylogenetic diversity measures.

All the rarefaction and extrapolation estimators proposed in this paper are featured in the online freeware application iNEXT-pd (iNterpolation/EXTrapolation for phylogenetic diversity) http://chao.stat.nthu.edu.tw/blog/software-download/.

Acknowledgements

The authors thank the Editor (Robert B. O'Hara), Lou Jost, Frederick Matsen and one anonymous reviewer for carefully reading an earlier version and providing very helpful suggestions and comments. This research is supported by Taiwan National Science Council under Contract 100-2118-M007-006-MY3. C-H Chiu is supported by a post-doctoral fellowship, National Tsing Hua University. T. C. Hsieh is supported by a post-doctoral fellowship, Taiwan National Science Council. D. A. Nipperess is supported by the Australian Research Council (DP1095200). Thomas Davis and David Nipperess would like to thank Emily Cave and Marina Tokarski for assistance with bird data collection. D.P. Faith thanks members of bioGENESIS and GEO BON for helpful discussions.

Data accessibility

All data used in this manuscript are present in the manuscript and its supporting information.

References

Biedermann, L., Zeitz, J., Mwinyi, J., Sutter-Minder, E., Rehman, A., Ott, S.J. *et al.* (2013) Smoking cessation induces profound changes in the composition of the intestinal microbiota in humans. *PLoS ONE*, **8**, e59260.

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

Cardoso, P., Rigal, F., Borges, P.A.V. & Carvalho, J.C. (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., Ackerly, D.D. & Kozak, K.H. (2012) Integrating ecology and phylogenetics: the footprint of history in modern-day communities 1. *Ecology*, **93**, 1–3.

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

Chao, A. (2005) Species estimation and applications. *Encyclopedia of Statistical Sciences* (eds S. Kotz, N. Balakrishnan, C.B. Read & B. Vidakovic), pp. 7907–7916. Wiley, New York.

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

Chao, A., Chiu, C.-H. & Jost, L. (2014) Unifying species diversity, phylogenetic diversity, functional diversity and related similarity/differentiation measures through Hill numbers. *The Annual Review of Ecology, Evolution, and Systematics*.

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

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

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

Chiu, C.-H., Wang, Y.T., Walther, B.A. & Chao, A. (2014) An improved non-parametric lower bound of species richness via a modified Good-Turing frequency formula. *Biometrics*, doi: 10.1111/biom.12200.

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.

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

Faith, D.P. (1992b) Systematics and conservation: on predicting the feature diversity of subsets of taxa. *Cladistics*, **8**, 361–373.

Faith, D.P. (2013) Biodiversity and evolutionary history: useful extensions of the PD phylogenetic diversity assessment framework. *Annals of the New York Academy of Sciences*, **1289**, 69–89.

Good, I.J. (1953) The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

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.

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.

Hill, M. (1973) Diversity and evenness: a unifying notation and its consequences. *Ecology*, **54**, 427–432.

Jetz, W., Thomas, G.H., Joy, J.B., Hartmann, K. & Mooers, A.O. (2012) The global diversity of birds in space and time. *Nature*, **491**, 444–448.

Kembel, S.W., Jones, E., Kline, J., Northcutt, D., Stenson, J., Womack, A.M. *et al.* (2012) Architectural design influences the diversity and structure of the built environment microbiome. *The ISME Journal*, **6**, 1469–1479.

Lauber, C.L., Hamady, M., Knight, R. & Fierer, N. (2009) Pyrosequencing-based assessment of soil pH as a predictor of soil bacterial community structure at the continental scale. *Applied and Environmental Microbiology*, **75**, 5111.

Magurran, A.E. & McGill, B.J. (eds) (2011) *Biological Diversity: Frontiers in Measurement and Assessment*. Oxford University Press, Oxford.

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

Paulson, J.N., Stine, O.C., Bravo, H.C. & Pop, M. (2013) Differential abundance analysis for microbial marker-gene surveys. *Nature Methods*, **10**, 1200–1202.

Pielou, E.C. (1975) *Ecological Diversity*. Wiley, New York.

Schlaeppi, K., Dombrowski, N., Oter, R.G., van Themaat, E.V.L. & Schulze-Lefert, P. (2014) Quantitative divergence of the bacterial root microbiota in Arabidopsis thaliana relatives. *Proceedings of National Academy Science, USA*, **111**, 585–592.

Williams, S.E., Shoo, L.P., Henriod, R. & Pearson, R.G. (2010) Elevational gradients in species abundance, assemblage structure and energy use of rainforest birds in the Australian Wet Tropics bioregion. *Australian Ecology*, **35**, 650–664.

Received 11 April 2014; accepted 22 July 2014
Handling Editor: Robert B. O'Hara

Supporting Information

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

Appendix S1. Derivation of the formulae for *PD* accumulation curve and *PD* rarefaction.

Appendix S2. A bootstrap method to obtain approximate unconditional variances of *PD* estimators and associated confidence intervals.

Appendix S3. The performance of the proposed Chao1-*PD* estimator.

Appendix S4. Derivation of the formula for *PD* extrapolation.

Appendix S5. Simulation results for rarefaction and extrapolation estimators.

Appendix S6. The phylogenetic tree and statistics for bird data in New South Wales.

Appendix S7. Derivation of the formulae for *PD* rarefaction and extrapolation (incidence data).