Syst. Biol. 66(1):100–111, 2017

DOI:10.1093/sysbio/syw073
Advance Access publication August 26, 2016

Rarefaction and Extrapolation: Making Fair Comparison of Abundance-Sensitive Phylogenetic Diversity among Multiple Assemblages

T. C. Hsieh and Anne Chao*

Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan;
**Correspondence to be sent to:Institute of Statistics, National Tsing Hua University, Hsin-Chu 30043, Taiwan; E-mail: chao@stat.nthu.edu.tw.*

Received 18 August 2015; reviews returned 22 August 2016; accepted 22 August 2016
Associate Editor: Dan Faith

Abstract.—Measures of phylogenetic diversity are basic tools in many studies of systematic biology. Faith's PD (sum of branch lengths of a phylogenetic tree connecting all focal species) is the most widely used phylogenetic measure. Like species richness, Faith's PD based on sampling data is highly dependent on sample size and sample completeness. The sample-size- and sample-coverage-based integration of rarefaction and extrapolation of Faith's PD was recently developed to make fair comparison across multiple assemblages. However, species abundances are not considered in Faith's PD. Based on the framework of Hill numbers, Faith's PD was generalized to a class of phylogenetic diversity measures that incorporates species abundances. In this article, we develop both theoretical formulae and analytic estimators for seamless rarefaction and extrapolation for this class of abundance-sensitive phylogenetic measures, which includes simple transformations of phylogenetic entropy and of quadratic entropy. This work generalizes the previous rarefaction/extrapolation model of Faith's PD to incorporate species abundance, and also extends the previous rarefaction/extrapolation model of Hill numbers to include phylogenetic differences among species. Thus a unified approach to assessing and comparing species/taxonomic diversity and phylogenetic diversity can be established. A bootstrap method is suggested for constructing confidence intervals around the phylogenetic diversity, facilitating the comparison of multiple assemblages. Our formulation and estimators can be extended to incidence data collected from multiple sampling units. We also illustrate the formulae and estimators using bacterial sequence data from the human distal esophagus and phyllostomid bat data from three habitats. [Extrapolation; diversity; Hill numbers; interpolation; phylogenetic diversity; prediction; rarefaction; sample completeness; sample coverage.]

Many studies in systematic biology require robust and sensible measures for quantifying and comparing phylogenetic diversity. An enormous number of phylogenetic diversity measures have been proposed (e.g., see Faith 1992; Warwick and Clarke 1995; McPeek and Miller 1996; Webb et al. 2006; Cavender-Bares et al. 2012 among others). The most widely used phylogenetic metric is Faith's PD (Faith 1992) which is defined as the sum of the branch lengths of a phylogenetic tree connecting all species in the target assemblage, but species abundances are not considered in Faith's PD. For some applications, the mere presence or absence of a species is all that can be determined from the available data, or all that is needed for the question at hand. In those cases, Faith's PD is a good measure of phylogenetic diversity. However, the ecological interactions in most communities depend on both the species abundances and the evolutionary histories of focal species (e.g., see Cadotte et al. 2010). When species abundances are available, there are advantages to incorporating abundance information into phylogenetic diversity measures. For example, some human impacts can result in the phylogenetic simplification of an ecosystem, reducing the population shares of phylogenetically distinct species relative to typical species (e.g., Helmus et al. 2010). An abundance-sensitive measure can reflect this effect and improve our ability to understand the impact of evolutionary history on ecological processes and patterns. As complements of Faith's PD, abundance-sensitive phylogenetic measures can provide additional useful information (e.g., Kembel et al. 2011; McCoy and Matsen 2013). See *Applications* for further justification of abundance-based measures for microbial sequence data.

Most previous abundance-based phylogenetic measures are generalizations of classic abundance-based species/taxonomic diversity indices. Rao's quadratic entropy (Rao 1982), a generalization of the classic Gini–Simpson index, was the first diversity measure to account for both phylogeny and species abundances. It can be interpreted as the average phylogenetic distance between any two randomly selected individuals from the assemblage. Allen et al. (2009) extended the classic Shannon entropy to phylogenetic entropy, which also incorporated phylogenetic distances among species.

An essential property that captures biologists' intuitive notion of diversity is the replication principle (MacArthur 1965; Hill 1973). This property requires that if we have N equally diverse, equally large assemblages with no species in common, the diversity of the pooled assemblage must be N times the diversity of a single assemblage. Classic diversity measures, such as Shannon entropy and the Gini–Simpson index, do *not* obey this principle and can lead to inconsistent or counter-intuitive interpretations if these measures are used to quantify diversity (Jost 2006, 2007). Consequently, their phylogenetic generalizations do not satisfy the replication principle either. Their generalizations will therefore have the same interpretational problems as their parent measures; see Chao et al. (2010, their Supplementary Material) for examples.

For classic abundance-based diversity measures, MacArthur (1965) solved the previously mentioned problem by simple transformations; Shannon entropy

can be converted by taking its exponential, and the Gini–Simpson index can be converted by taking the inverse of its complement. These converted measures then satisfy the replication principle. Hill (1973) integrated species richness and these two converted measures into a continuum of diversity measures called Hill numbers, or the effective number of species, parameterized by a diversity order q. Here, "effective number" means the number of equally abundant and equally distinct species that are needed to give the same value of the original measure. Hill numbers include three widely used measures: species richness ($q=0$), Shannon diversity (the exponential of Shannon entropy, $q=1$), and Simpson diversity (the inverse of Simpson concentration, $q=2$), all in units of "species equivalents" (Hill 1973). Hill numbers obey the replication principle and behave like "species richness"; they also yield self-consistent assessment, have intuitively interpretable magnitudes, and can be meaningfully decomposed (Chao et al. 2014a). Thus, Hill numbers resolve many of the interpretational problems caused by classic diversity indices.

Chao et al. (2010) extended Hill numbers to a class of phylogenetic diversity measures that satisfy the replication principle. This class of phylogenetic measures include Faith's PD ($q=0$), a simple transformation of phylogenetic entropy ($q=1$), and a simple transformation of quadratic entropy ($q=2$). Chao et al. (2010) measures were subsequently extended by Faith and Richards (2012) and Faith (2013).

Based on sampling data, both Hill numbers (including species richness) and their phylogenetic generalization (including Faith's PD) depend on sample size and inventory completeness (Chao et al. 2014b). Thus, standardization is needed to make fair comparison and assessment of diversities across multiple assemblages. The traditional standardization to compare species richnesses of different assemblages is to use rarefaction to down-sample the larger samples until they are the same size as the smallest sample (e.g., Gotelli and Colwell 2001). Ecologists then compare the richnesses of these equally large samples, but 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 for species richness that can be rarefied to smaller sample sizes or extrapolated to a larger sample size, guided by an estimate of asymptotic richness. Chao and 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 with traditional rarefaction/extrapolation to equal sample sizes. They developed coverage-based rarefaction/extrapolation methodology for species richness to implement this approach. Chao et al. (2014b) extended Colwell et al. (2012) and Chao and Jost (2012) to Hill numbers and proposed the sample-size- and coverage-based integration of rarefaction and extrapolation of Hill numbers as a general framework for estimating abundance-based taxonomic diversity and for making statistical inferences based on these estimates.

Compared with species diversity, statistical estimation and standardization of phylogenetic diversity have rarely been explored. Ricotta et al. (2012) developed the sample-size-based rarefaction formula for quadratic entropy. Nipperess and Matsen (2013) derived the exact analytic formula for the mean and (conditional) variance of Faith's PD under sample-size-based rarefaction. Chao et al. (2015) subsequently developed an integrated rarefaction and extrapolation sampling curve for Faith's PD.

Until now, the rarefaction/extrapolation sampling curves for abundance-sensitive phylogenetic diversity have not been developed, because the extension is not conceptually and technically direct. To fill in this gap, we derive here for the first time both theoretical formulae and analytic estimators for seamless rarefaction/extrapolation for Chao et al. (2010) phylogenetic measures. This work generalizes the previous rarefaction/extrapolation models of Faith's PD to incorporate abundances, and also extends rarefaction/extrapolation of Hill numbers to include phylogenetic differences among species. Therefore, a unified sampling framework based on rarefaction/extrapolation for analyzing biodiversity data can be established. A bootstrap method is also suggested for constructing confidence intervals around the phylogenetic measures, facilitating the comparison of multiple assemblages. Bacterial sequence data in the human distal esophagus are used for illustration.

Review: A Class of Phylogenetic Diversity Measures of Order q

Let S denote the number of species in an assemblage, and let p_i represent the relative abundance of species i, $i=1, 2, \ldots, S$, such that $\sum_{i=1}^{S} p_i = 1$. Hill numbers of order q are defined as the following function of relative abundances $\{p_1, p_2, \ldots, p_S\}$:

$$ {}^{q}D = \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)}, \quad q \geqslant 0, q \neq 1. \tag{1a} $$

If ${}^{q}D = k$, then the diversity of order q of the actual assemblage is the same as that of an assemblage with k equally abundant and equally distinct species. This measure is undefined for $q=1$, but its limit as q tends to unity exists:

$$ {}^{1}D = \lim_{q \to 1} {}^{q}D = \exp\left(-\sum_{i=1}^{S} p_i \log p_i \right). \tag{1b} $$

To formulate phylogenetic diversity, we assume that all species of this assemblage are connected by a rooted ultrametric or nonultrametric phylogenetic tree, with the S species as tip nodes. Throughout the article, all

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 sampling data. If this reference point is chosen to be the root, then the phylogenetic diversity of order 0 reduces to Faith (1992) definition; see *Discussion* for the choice of a reference point. Assume that there are B branch segments in the tree, $B \geqslant S$, for the given reference point on the main trunk. Let L_i denote the length of the ith branch, and a_i denote the total relative abundance descended from the ith branch, $i=1, 2, \ldots, B$. With this definition of a_i, the set of species relative abundances $(p_1, p_2, \ldots, p_S)$ is expanded to the set of branch relative abundances $\{a_i, i=1, 2, \ldots, B\}$ with $(p_1, p_2, \ldots, p_S)$ as the first S elements. We refer to a_i as the branch relative abundance of the ith branch segment, although $\sum_{i=1}^{B} a_i$ typically is greater than 1.

Under our formulation, Faith's PD is $\sum_{i=1}^{B} L_i$, the total branch length from the given reference point as described earlier. Rao's quadratic entropy Q can be expressed as (Rao 1982):

$$Q = \sum_{i,j} d_{ij} p_i p_j, \tag{2a}$$

where d_{ij} denotes the phylogenetic distance (in years since divergence, number of DNA base changes, or other metrics) between species i and j, and p_i and p_j denote the relative abundance of species i and j. Rao's Q represents a phylogenetic generalization of the Gini–Simpson index because in the special case of no phylogenetic structure (all species are equally related to one another), $d_{ii}=0$ and $d_{ij}=1$ $(i \neq j)$, it reduces to the Gini–Simpson index. The formula for the phylogenetic entropy H_P is (Allen et al. 2009):

$$H_P = -\sum_{i=1}^{B} L_i a_i \log a_i \tag{2b}$$

Let $\bar{T} = \sum_{i=1}^{B} L_i a_i$ denote the mean branch length (weighted by branch relative abundance); it can also be interpreted as the average lineage length where a "lineage length" is the total path length from a reference point to each of the terminal branch tips (Fig. 1 of Chao et al. 2010). For an ultrametric tree, $\bar{T}$ simply reduces to the tree depth. Chao et al. (2010) derived the following phylogenetic diversity of order q:

$${}^{q}\mathrm{PD} = {}^{q}\mathrm{PD}(\infty) = \left(\sum_{i=1}^{B} L_i \left(\frac{a_i}{\bar{T}} \right)^{q} \right)^{1/(1-q)}, \quad q \neq 1; \tag{3a}$$

$$\begin{aligned} {}^{1}\mathrm{PD} &= {}^{1}\mathrm{PD}(\infty) = \lim_{q \to 1} {}^{q}\mathrm{PD}(\infty) \\ &= \exp\left(-\sum_{i=1}^{B} L_i \frac{a_i}{\bar{T}} \log \frac{a_i}{\bar{T}} \right), \quad q=1. \end{aligned} \tag{3b}$$

These measures also represent the asymptotic diversities at a sample size of infinity, so we use the notation ${}^{q}\mathrm{PD}(\infty)$ and ${}^{q}\mathrm{PD}$ interchangeably for clearer presentation in later derivation steps. The phylogenetic diversity ${}^{q}\mathrm{PD}$ is interpreted as the effective total branch length in the assemblage from the given reference point of the phylogenetic tree. Thus, the ${}^{q}\mathrm{PD}$ measure for two assemblages with different values of $\bar{T}$ can still be meaningfully compared. If there are no internal nodes and all are standardized to be unity, then ${}^{q}\mathrm{PD}$ reduces to the Hill numbers ${}^{q}\mathrm{D}$ defined in Equation (1a).

The diversity order q determines the measures' emphasis on rare or common branches. When $q=0$, ${}^{0}\mathrm{PD}$ is Faith's PD, the actual total branch length and branch abundances are not considered. When $q=1$, the measure ${}^{1}\mathrm{PD}$ is interpreted as the effective total branch length, which weights each branch in proportional to its abundance; it is a monotonic transformation of the phylogenetic entropy H_P: ${}^{1}\mathrm{PD} = \bar{T}\exp(H_P/\bar{T})$. When $q=2$, ${}^{2}\mathrm{PD}$ is interpreted as the effective total branch length, which puts more weight on abundant/dominant branches and discounts rare branches; it is a monotonic transformation of Rao's quadratic entropy Q: ${}^{2}\mathrm{PD} = \bar{T}^2/(\bar{T}-Q)$.

Chao et al. (2010, 2014a) and Chiu et al. (2014) proved the measure ${}^{q}\mathrm{PD}$ for $q \geqslant 0$ satisfies the phylogenetic version of the replication principle based on species relative abundance data. In Online Appendix A (available on Dryad at http://dx.doi.org/10.5061/dryad.qk58h), we extend it to any measure of species importance including raw abundances: Assume there are N completely phylogenetically distinct assemblages (no shared branches across assemblages, though branches within an assemblage may be shared among species); see Fig. 1 of Chiu et al. (2014) for examples. Suppose these assemblages have identical qth order phylogenetic diversities X and identical length-weighted total abundances (the latter condition reduces to "identical mean branch lengths" for species relative abundance data), then the pooled assemblage must have a phylogenetic diversity $N \times X$ for the same order q. Since the replication principle is an essential property for our phylogenetic measures, we discuss this property with proof details in Online Appendix A, available on Dryad.

Material and Methods

A Reference Sample of Size n

We assume an empirical sample of n individuals is taken with replacement from the assemblage with species relative abundances $\{p_1, p_2, \ldots, p_S\}$ and branch relative abundances $\{a_i, i=1, 2, \ldots, B\}$ for a given phylogenetic tree of S species. We follow Colwell et al. (2012) to call this observed sample (including the observed tree which is spanned by the observed species) as the *reference sample*. Let X_i denote the observed

abundance of the ith species in the reference sample; only species with $X_i > 0$ is observed, $\sum_{i=1}^{S} X_i = n$. We expand the set of sample species abundances/frequencies $\{X_1, X_2, ..., X_S\}$ to a larger set of sample branch/node abundances $\{X_i^*, i = 1, 2, ..., B\}$ by defining X_i^* as the sum of the sample abundances for those species descended from the ith branch.

Define g_k as the sum of branch lengths for those branches with sample branch abundance k, that is,

$$g_k = \sum_{i=1}^{B} L_i I(X_i^* = k), \quad k = 0, 1, \ldots n, \tag{4a}$$

where $I(\cdot)$ is an indicator function that equals 1 when true and 0 otherwise. Thus, g_0 represents the total length of branches that are not detected in the observed tree; g_0 is unknown but $\{g_1, g_2, ...\}$ can be computed from the reference sample. In particular, g_1 denotes the total branch length of those singletons, and g_2 denotes the total branch length of those doubletons in the branch abundance set $\{X_i^*, i = 1, 2, ..., B\}$. Equivalently, g_1 (g_2) denotes the total branch length of those singletons (doubletons) in the sample branch abundance set of the observed tree.

Based on Equation (3a), the observed phylogenetic diversity in the reference sample can be expressed as

$$\begin{aligned} {}^{q}\mathrm{PD}_{\mathrm{obs}} &= \left[\sum_{i=1}^{B} L_i \left(\frac{X_i^*/n}{\bar{T}_{\mathrm{obs}}} \right)^q \right]^{1/(1-q)} \\ &= \left[\sum_{k=1}^{n} \left(\frac{k/n}{\bar{T}_{\mathrm{obs}}} \right)^q \times g_k \right]^{1/(1-q)}. \end{aligned} \tag{4b}$$

Here for a nonultrametric tree, $\bar{T}_{\mathrm{obs}} = \sum_{i=1}^{B} L_i (X_i^*/n)$ represents the observed or empirical mean branch length in the reference sample. For an ultrametric tree, $\bar{T}_{\mathrm{obs}}$ automatically reduces to the tree depth.

^{q}PD *Accumulation Curve*

To derive the theoretical formula for the expected phylogenetic diversity as a function of sample size $m = 1, 2, \ldots$, we assume a hypothetical sample of m individuals is taken from the entire assemblage. As we defined $\{X_i^*, i = 1, 2, ..., B\}$ and $\{g_k, k = 0, 1, \ldots, n\}$ for the reference sample of size n in the preceding subsection, we can similarly define $\{X_i^*(m), i = 1, 2, ..., B\}$ and $\{g_k(m), k = 0, 1, \ldots, m\}$ for a hypothetical sample of size m. That is, $X_i^*(m)$ represents the sample abundance of branch i; $g_k(m)$ represents the total length of those branches with abundance k in a hypothetical sample of size m, that is,

$$g_k(m) = \sum_{i=1}^{B} L_i I(X_i^*(m) = k), \quad k = 0, 1, \ldots, m. \tag{5a}$$

Thus, we extend the definition of X_i^* and g_k for the reference sample size n to any hypothetical sample size m. Under the commonly used multinomial sampling assumption, $X_i^*(m)$, $i = 1, 2, \ldots, B$, follows a binomial distribution with sample size m and probability a_i (branch relative abundance), we have

$$E[g_k(m)] = \sum_{i=1}^{B} L_i \binom{m}{k} a_i^k (1 - a_i)^{m-k}, \quad k = 0, 1, ..., m. \tag{5b}$$

In the special case of $k = 0$, the above equation reduces to $E[g_0(m)] = \sum_{i=1}^{B} L_i (1 - a_i)^m$, the expected sum of undetected branch lengths in a sample of size m; we also have $\sum_{k=0}^{m} E[g_k(m)] = \sum_{i=1}^{B} L_i$, which is Faith's PD.

Define ${}^{q}\mathrm{PD}(m)$ for any positive integer m as the phylogenetic diversity based on the expected distribution of sample branch abundances (equivalently, the distribution of $g_k(m)$, $k = 0, 1, \ldots, m$) in a hypothetical sample of size m. In this hypothetical sample, any sample branch relative abundance is in a form of k/m, $k = 0, 1, \ldots, m$. The expected total length of those branches with sample branch relative abundance k/m is $E[g_k(m)]$. Using Equation (5b), we can evaluate the expected mean length as $\sum_{k=1}^{m} (k/m) \times E[g_k(m)] = \sum_{i=1}^{B} L_i a_i$, which is $\bar{T}$ defined earlier. Substituting L_i and a_i in Equations (3a) and (3b) by $E[g_k(m)]$ and k/m, we obtain the following expected diversities:

$${}^{q}\mathrm{PD}(m) = \left[\sum_{k=1}^{m} E[g_k(m)] \times \left(\frac{k/m}{\bar{T}} \right)^q \right]^{1/(1-q)}, \quad q \neq 1; \tag{6a}$$

$${}^{1}\mathrm{PD}(m) = \exp\left[-\sum_{k=1}^{m} E[g_k(m)] \times \frac{k/m}{\bar{T}} \log \frac{k/m}{\bar{T}} \right]. \tag{6b}$$

The above theoretical expressions are valid for any nonnegative integer $m = 1, 2, \ldots$.

The plot of ${}^{q}\mathrm{PD}(m)$ with respect to the sample size m is referred to as a sample-size-based ^{q}PD accumulation curve. As sample size m tends to infinity, ${}^{q}\mathrm{PD}(m)$ approaches the asymptotic value ${}^{q}\mathrm{PD}(\infty) = {}^{q}\mathrm{PD}$. Equation (6a) for $q = 0$ reduces to the formula for the expected Faith's PD in a hypothetical sample of size m (Chao et al. 2015):

$$\begin{aligned} {}^{0}\mathrm{PD}(m) &= \sum_{k=1}^{m} E[g_k(m)] = \sum_{i=1}^{B} L_i - E[g_0(m)] \\ &= \sum_{i=1}^{B} L_i [1 - (1 - a_i)^m]. \end{aligned}$$

When there are no internal nodes and all branches are equally distinct with branch lengths of unity (i.e., branch lengths are normalized to unity), the ^{q}PD accumulation curve reduces to the corresponding Hill

number accumulation curve derived in Chao et al. (2014b). In this special case, the curve for $q=0$ reduces to the classic species accumulation curve.

Chao and Jost (2012) proposed standardizing samples by sample completeness as measured by sample coverage (or simply coverage), a concept originally developed by A. Turing and I.J. Good in their famous cryptographic analysis during World War II (Good 1953, 2000). They defined the sample coverage of a given sample 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 hypothetical sample of size m. The plot of ${}^{q}\mathrm{PD}(m)$ as a function of $C(m)$ is the coverage-based ${}^{q}\mathrm{PD}$ accumulation curve. As $C(m)$ tends to unity (complete coverage), the diversity approaches the asymptotic value ${}^{q}\mathrm{PD}(\infty)={}^{q}\mathrm{PD}$.

In the following subsections, we demonstrate how to estimate the two types (sample-size- or coverage-based) of ${}^{q}\mathrm{PD}$ accumulation curve based on the reference sample of size n. Rarefaction refers to the estimation of ${}^{q}\mathrm{PD}(m)$ for a hypothetical sample of size $m\leqslant n$, and extrapolation refers to the estimation of ${}^{q}\mathrm{PD}(m)$ for $m=n+m^{*}>n$ (Table 1).

Rarefaction of ${}^{q}\mathrm{PD}$ Measure

Our approach to deriving rarefaction formulae is based on Equations (6a) and (6b) by substituting $E[g_k(m)]$ with its estimator. From Chao et al. (2014b, their Equation 8), an unbiased estimator for $E[g_k(m)]$, $m\leqslant n$, $k\geqslant 1$, can be similarly derived as:

$$\hat{g}_k(m)=\sum_{\substack{i=1\\ k\leqslant X_i^{*}\leqslant n-m+k}}^{B} L_i\frac{\binom{X_i^{*}}{k}\binom{n-X_i^{*}}{m-k}}{\binom{n}{m}} =\sum_{k\leqslant j\leqslant n-m+k}\frac{\binom{j}{k}\binom{n-j}{m-k}}{\binom{n}{m}}g_j. \tag{7}$$

We thus obtain the following analytic estimators of the expected diversity of a rarefied sample of size m ($m\leqslant n$):

$${}^{q}\widehat{\mathrm{PD}}(m)=\left[\sum_{k=1}^{m}\left(\frac{k/m}{\bar{T}_{\mathrm{obs}}}\right)^{q}\times\hat{g}_k(m)\right]^{1/(1-q)},\quad q\neq 1; \tag{8a}$$

$${}^{1}\widehat{\mathrm{PD}}(m)=\exp\left[-\sum_{k=1}^{m}\left(\frac{k/m}{\bar{T}_{\mathrm{obs}}}\log\frac{k/m}{\bar{T}_{\mathrm{obs}}}\right)\times\hat{g}_k(m)\right],\quad q=1. \tag{8b}$$

Equation (8a) is an unbiased estimator for order $q=0$ and nearly unbiased for ${}^{q}\mathrm{PD}(m)$ for any order $q>0$ because they are obtained by taking nonlinear functions of an unbiased estimator. "Nearly unbiased" means that its bias tends to zero as the reference sample size n becomes large.

Extrapolation of ${}^{q}\mathrm{PD}$ Measure

Based on a reference sample of size n, the extrapolation of ${}^{q}\mathrm{PD}$ of order q is a prediction of ${}^{q}\mathrm{PD}(n+m^{*})$ for a sample of size $m=n+m^{*}$, $m^{*}>0$. Unlike rarefaction, unbiased estimators for $E[g_k(n+m^{*})]$ do not exist, so more sophisticated methods based on estimated asymptotes are needed for extrapolation; see Online Appendix B, available on Dryad, for all derivation details. The general extrapolation formula for $q\geqslant 0$ is based on the following formula, guided by an asymptotic diversity estimator ${}^{q}\widehat{\mathrm{PD}}(\infty)$:

$${}^{q}\widehat{\mathrm{PD}}(n+m^{*})={}^{q}\mathrm{PD}_{\mathrm{obs}}+[{}^{q}\widehat{\mathrm{PD}}(\infty)-{}^{q}\mathrm{PD}_{\mathrm{obs}}]\,[1-(1-{}^{q}\hat{\beta})^{m^{*}}],q\geqslant 0, \tag{9a}$$

and

$${}^{q}\hat{\beta}=[{}^{q}\mathrm{PD}_{\mathrm{obs}}-{}^{q}\widehat{\mathrm{PD}}(n-1)]/[{}^{q}\widehat{\mathrm{PD}}(\infty)-{}^{q}\widehat{\mathrm{PD}}(n-1)],q\geqslant 0, \tag{9b}$$

where ${}^{q}\widehat{\mathrm{PD}}(n-1)$ is a diversity estimator for a rarefied sample of size $n-1$ (given in Equations 8a and 8b). All formulae for the special cases of $q=0$, $q=1$, $q=2$, and an integer $q>2$ are shown in Table 1.

The extrapolation formula for $q=0$ (Faith's PD) is guided by a Chao1-PD estimator (Chao et al. 2015) for the true Faith's PD which includes the observed total branch length in the reference sample and an estimator of g_0; see Equation (B.2) of Online Appendix B, available on Dryad, for the Chao1–PD estimator ${}^{0}\widehat{\mathrm{PD}}(\infty)$ and the corresponding estimator $\hat{g}_0$. The estimator ${}^{0}\hat{\beta}$ in Equation (9b) can be expressed as

$${}^{0}\hat{\beta}=g_1/(n\hat{g}_0+g_1). \tag{9c}$$

Combining Equations (9a) and (9c), we can obtain the prediction formula for $q=0$ (Chao et al. 2015):

$${}^{0}\widehat{\mathrm{PD}}(n+m^{*})={}^{0}\mathrm{PD}_{\mathrm{obs}}+\hat{g}_0\left[1-(1-\frac{g_1}{n\hat{g}_0})^{m^{*}}\right].$$

For a short-range prediction (e.g., m^{*} is much less than n), an approximation formula is

$${}^{0}\widehat{\mathrm{PD}}(n+m^{*})\approx{}^{0}\mathrm{PD}_{\mathrm{obs}}+(g_1/n)m^{*}.$$

In this case, the right-most expression of the above formula reveals that the extrapolation is independent of the choice of $\hat{g}_0$. Consequently, our extrapolation formula for $q=0$ for a short-range extrapolation is very robust and reliable even though the Chao1-PD estimator is a lower bound. The same conclusion extends to any order $q\geqslant 0$.

The extrapolation formula for the case $q=1$ is guided by an estimator of the phylogenetic entropy because of the relationship ${}^{1}\mathrm{PD}(\infty)=\bar{T}\exp(H_P/\bar{T})$. Even for ordinary Shannon entropy, the estimation is surprisingly

TABLE 1. The theoretical formulae and analytic estimators for rarefaction and extrapolation of phylogenetic diversity of order $q=0$ (first row of equations), $q=1$ (second row), $q=2$ (third row) and any integer order $q>2$ (fourth row), given a reference sample[a] with the observed ${}^{q}\mathrm{PD}_{\mathrm{obs}}$ and estimated sample coverage $\hat{C}(n)$

Theoretical formula[b] (for all $m>0$)	Interpolation estimator[c] (for $m<n$)	Extrapolation estimator[d] (for a sample of size $n+m^*$)
${}^{0}\mathrm{PD}(m)=\sum_{k=1}^{m} E[g_k(m)]$	${}^{0}\widehat{\mathrm{PD}}(m)=\sum_{k=1}^{m}\hat{g}_k(m)$ (unbiased)	${}^{0}\widehat{\mathrm{PD}}(n+m^*)$ $={}^{0}\mathrm{PD}_{\mathrm{obs}}+[{}^{0}\widehat{\mathrm{PD}}(\infty)-{}^{0}\mathrm{PD}_{\mathrm{obs}}][1-(1-{}^{0}\hat{\beta})^{m^*}]$ (reliable if $m^*<n$)
${}^{1}\mathrm{PD}(m)$ $=\exp\left(-\sum_{k=1}^{m}\left(\frac{k/m}{\bar{T}}\log\frac{k/m}{\bar{T}}\right)\times E[g_k(m)]\right)$	${}^{1}\widehat{\mathrm{PD}}(m)$ $=\exp\left[-\sum_{k=1}^{m}\left(\frac{k/m}{\bar{T}_{\mathrm{obs}}}\log\frac{k/m}{\bar{T}_{\mathrm{obs}}}\right)\times\hat{g}_k(m)\right]$ (nearly unbiased)	${}^{1}\widehat{\mathrm{PD}}(n+m^*)$ $={}^{1}\mathrm{PD}_{\mathrm{obs}}+[{}^{1}\widehat{\mathrm{PD}}(\infty)-{}^{1}\mathrm{PD}_{\mathrm{obs}}][1-(1-{}^{1}\hat{\beta})^{m^*}]$ (nearly unbiased)
${}^{2}\mathrm{PD}(m)$ $=\frac{1}{\sum_{k=1}^{m}\left(\frac{k/m}{\bar{T}}\right)^{2}\times E[g_k(m)]}$	${}^{2}\widehat{\mathrm{PD}}(m)=\frac{1}{\sum_{k=1}^{m}\left(\frac{k/m}{\bar{T}_{\mathrm{obs}}}\right)^{2}\times\hat{g}_k(m)}$ (nearly unbiased)	${}^{2}\widehat{\mathrm{PD}}(n+m^*)$ $=\frac{1}{\sum_{i=1}^{B}\frac{L_i}{\bar{T}_{\mathrm{obs}}^{2}}\times\left(\frac{1}{n+m^*}\frac{X_i^*}{n}+\frac{(n+m^*-1)}{n+m^*}\frac{X_i^*(X_i^*-1)}{n(n-1)}\right)}$ (nearly unbiased)
${}^{q}\mathrm{PD}(m)$ $=\left(\sum_{k=1}^{m}\left(\frac{k/m}{\bar{T}}\right)^{q}\times E[g_k(m)]\right)^{\frac{1}{1-q}}$	${}^{q}\widehat{\mathrm{PD}}(m)=\left[\sum_{k=1}^{m}\left(\frac{k/m}{\bar{T}_{\mathrm{obs}}}\right)^{q}\times\hat{g}_k(m)\right]^{1/(1-q)}$ (nearly unbiased)	${}^{q}\widehat{\mathrm{PD}}(n+m^*)$ $=\left[\sum_{i=1}^{B}\frac{L_i}{\bar{T}_{\mathrm{obs}}^{q}}\times\left(\sum_{j=1}^{q}\psi(q,j)\frac{(n+m^*)^{(j)}}{(n+m^*)^{q}}\frac{(X_i^*)^{(j)}}{n^{(j)}}\right)\right]^{1/(1-q)}$ (nearly unbiased)
[e]Expected coverage of sample size m: $C(m)=1-\sum_{i=1}^{S}p_i(1-p_i)^{m}$	$\hat{C}(m)=1-\sum_{i=1}^{S}\frac{X_i}{n}\frac{\binom{n-X_i}{m}}{\binom{n-1}{m}}$ (unbiased)	$\hat{C}(n+m^*)=1-\frac{f_1}{n}\left[\frac{(n-1)f_1}{(n-1)f_1+2f_2}\right]^{m^*+1}$ (reliable for $m^*<n$)

[a]The observed phylogenetic diversity of order q of the reference sample is given in Equation (4b). See Chao et al. (2014b, Table 1) for the coverage estimator of the reference sample.
[b]The term $g_k(m)$ is defined as the total branch lengths of those branches with sample branch abundance k in a hypothetical sample of size m; see Equation (5a). The formula for $E[g_k(m)]$ is given in Equation (5b).
[c]An unbiased estimator $\hat{g}_k(m)$ for $E[g_k(m)]$ for $m<n$ is given in Equation (7).
[d]For $q=0$ and 1, see Equation (9c) for the formula of ${}^{0}\hat{\beta}$, and see Online Appendix B, available on Dryad, for the formulae of ${}^{0}\widehat{\mathrm{PD}}(\infty)$, ${}^{1}\widehat{\mathrm{PD}}(\infty)$, and ${}^{1}\hat{\beta}$. For any integer $q\geqslant 2$, $x^{(j)}=x(x-1)\ldots(x-j+1)$ denotes the falling factorial, and $\psi(q,j)=$ the Stirling number of the second kind defined by the coefficient in the expansion $x^{q}=\sum_{j=1}^{q}\psi(q,j)x^{(j)}$.
[e]The last row gives equations for sample completeness as a function of sample size. It also gives the corresponding coverage estimators for rarefied samples and extrapolated samples for coverage-based rarefaction and extrapolation curves.

nontrivial as shown by Chao et al. (2013) who derived a nearly optimal estimator for Shannon entropy using the relationship between Shannon entropy and the slopes of the species accumulation curve. Similar arguments lead to an estimator $\hat{H}_P$ for the phylogenetic entropy (Online Appendix B, available on Dryad). Then the asymptote of ${}^{1}\mathrm{PD}$ can be estimated as ${}^{1}\widehat{\mathrm{PD}}(\infty)=\bar{T}_{\mathrm{obs}}\exp(\hat{H}_P/\bar{T}_{\mathrm{obs}})$. Based on this asymptotic estimator and Equation (9a), we obtain an extrapolated ${}^{1}\mathrm{PD}$ estimator at sample size $n+m^*$ (Table 1). For any integer $q\geqslant 2$, instead of using Equation (9a), we sugges using an exact extrapolation formula; see Table 1 for $q=2$ and Online Appendix B, available on Dryad, for a general integer q

A simulation study was conducted to investigate the performance of our analytic estimators for rarefaction, extrapolation, and asymptotes. We considered three scenarios for trees (one real tree for rockfish and two simulated trees generated from the R Package TreeSim available from CRAN) and two species abundance distributions. All results are reported in Online Appendix C, available on Dryad.

RESULTS

Note that all formulae in Table 1 for the special case of $q=0$ reduce to those derived in Chao et al. (2015) for Faith's PD. When there are no internal nodes,

and all branch lengths are standardized to be unity ($L_i = 1$ for all $i = 1, 2, ..., B$, and $\bar{T} = 1$), all formulae in Table 1 reduce to those for Hill numbers derived in Chao et al. (2014b). This work thus generalizes the previous rarefaction/extrapolation model of Faith's PD to incorporate species abundance, and also extends the rarefaction/extrapolation model of Hill numbers to include phylogeny among species. The unconditional variance estimator for the rarefied/extrapolated estimators along with the associated confidence interval can be computed by a bootstrap method; see Online Appendix S2, in Chao et al. (2015) for details. Generally, for any fixed sample size or any fixed degree of completeness in the comparison, 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 nonsignificance (Colwell et al. 2012) and thus are inconclusive.

Sample-Size-Based Rarefaction/Extrapolation

Our proposed sample-size-based sampling curve for ${}^{q}\mathrm{PD}$ includes the rarefaction part (which plots ${}^{q}\widehat{\mathrm{PD}}(m)$as a function of m, where $m \leqslant n$) and the extrapolation part (which plots ${}^{q}\widehat{\mathrm{PD}}(n+m^*)$ as a function of $n+m^*$, $m^* \geqslant 0$). The two parts join smoothly at the point of the reference sample (n, ${}^{q}\mathrm{PD}_{\mathrm{obs}}$), and the confidence intervals based on the bootstrap method also join smoothly. For the measure $q = 0$ (Faith's PD), the size can be extrapolated, at most, to no more than double the reference sample size (Chao et al. 2015). For the measures with $q = 1$ and $q = 2$ measures, if data are not sparse, the extrapolation can be reliably extended to infinity to attain the estimated asymptote (Online Appendix C, available on Dryad).

Coverage-Based Rarefaction/Extrapolation

For the expected coverage $C(m)$ of a sample of size m, Chao and Jost (2012) derived 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 $n+m^*$. Our proposed coverage-based sampling curve for ${}^{q}\mathrm{PD}$ includes rarefaction (which plots ${}^{q}\widehat{\mathrm{PD}}(m)$ with respect to $\hat{C}(m)$) and extrapolation (which plots ${}^{q}\widehat{\mathrm{PD}}(n+m^*)$ with respect to $\hat{C}(n+m^*)$) joining smoothly at the reference sample point ($\hat{C}(n)$, $\mathrm{PD}_{\mathrm{obs}}$). The confidence intervals based on the bootstrap method also join smoothly. For the measure $q = 0$ (Faith's PD), the maximum coverage for each sample is selected to be the coverage corresponding to the maximum sample size used in the sample-size-based rarefaction and extrapolation curve. For the measures with $q = 1$ and $q = 2$, if data are not sparse, the extrapolation can often be extended to the coverage of unity to attain the estimated asymptote.

Linking the Sample-size- and Coverage-Based Curves

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 subsection for an example.

Applications

We illustrate the proposed formulae and estimators by using the bacterial 16S rRNA sequence data from the human distal esophagus originally described by Pei et al. (2004). The data are also included in the software Mothur (Schloss et al. 2009), one of the most popular tools for analyzing microbial ecology data, as a demonstrative example. In Pei et al. (2004), four patients with normal esophageal histology and without any esophageal pathology were included. For our analysis, we selected the data of two patients (Subject C and Subject D) from Mothur's wiki page "Esophageal community analysis". The sequence processing and OTU clustering were performed based on a cut-off of 1% sequence difference. Under this clustering scheme, a total of 112 OTUs were found out of 472 sequences for the data of the two subjects (Online Appendix D, available on Dryad).

The numbers of sequences (reference sample size) for Subject C and Subject D are, respectively, 255 and 219, and the corresponding observed OTU richnesses are, respectively, 59 (with sample coverage 91.0%) and 69 (with sample coverage 81.2%). A phylogenetic tree (Online Appendix D, available on Dryad) of these OTUs was constructed by using greengenes database (DeSantis et al. 2006) and Mothur's analysis tool (Schloss et al. 2009). For illustrative purposes, we used the default distance (the proportion of bases that differ between each pair of sequences) and the neighbor-joining algorithm to construct the tree. We give the observed species diversity (Hill numbers) and phylogenetic diversity along with their estimated asymptotic values and 95% confidence intervals for $q = 0$, 1, and 2 for each subject (Table 2). Without loss of generality, we selected the root of all observed OTUs as our reference point. Although the root of the observed taxa varies with sampling data, we can easily transform all our estimates to those for any more basal reference point; see *Discussion* for the comparison of species and phylogenetic diversities as well as the relationship between two reference points.

In microbiology, there already are many bacterial 16S studies using PD rarefaction. For example, Lauber et al. (2009) compared soil bacterial communities by rarefying all samples to 1200 molecular sequences. Kembel et al. (2012) compared the microbial PD of different environments at a health-care facility by using

TABLE 2. The observed species diversity and phylogenetic diversity along with their estimated asymptotic values with their s.e. as well as 95% confidence intervals for $q=0$, 1, 2 for the bacterial sequence data from the human distal esophagus of two subjects (Pei et al. 2004)

Diversity	Subject	Order	Observed diversity	Estimated asymptote	Estimated s.e	95% confidence interval
Species Diversity (Hill numbers)	Subject C	$q=0$	59	77.8	10.28	(65.91, 110.24)
		$q=1$	32.83	38.95	2.78	(33.51, 44.39)
		$q=2$	21.77	23.71	2.03	(21.77, 27.68)
	Subject D	$q=0$	69	124.78	23.81	(94.02, 193.35)
		$q=1$	23.94	32.74	4.24	(24.42, 41.06)
		$q=2$	7.89	8.15	1.37	(7.89, 10.83)
Phylogenetic Diversity	Subject C	$q=0$	3.93	4.54	0.46	(3.93, 5.43)
		$q=1$	0.93	0.97	0.03	(0.93, 1.04)
		$q=2$	0.54	0.54	0.02	(0.54, 0.57)
	Subject D	$q=0$	4.52	8.66	1.35	(6.02, 11.3)
		$q=1$	0.72	0.77	0.04	(0.72, 0.85)
		$q=2$	0.46	0.46	0.01	(0.46, 0.48)

PD based on rarefied samples of 700 sequences. For humans, a number of illnesses are linked to a lowered microbial PD, probably reflecting a lowered "resilience" of the normal microbial community. These illnesses include Crohn disease, autism, pulmonary disease, chronic constipation, breast cancer risk, staphylococcus infection, obesity, and colon inflammation (Faith D., personal communication). Moreover, changes in relative abundance within such microbial communities may be important also, and thus highlight the need to make fair comparisons across communities via the rarefaction/extrapolation method for abundance-sensitive phylogenetic measures. These issues about resilience and so on extend to conventional "macrobial" communities as well. An additional analysis on phyllostomid bat data is used to further illustrate our proposed method (Online Appendix E, available on Dryad).

Following Chao et al. (2014b), we adopt the following three steps for comparing the phylogenetic diversity of the two subjects

Step 1: *Compare sample-size-based rarefaction/extrapolation curves up to a maximum size* (*Fig. 1a*).—To compare the diversity of the two subjects for equally large samples, we construct for each subject the sample-size-based rarefaction/extrapolation curves for ^{q}PD measures ($q=0$, 1, and 2) along with 95% confidence intervals up to a base sample size of 500 which is about double the reference sample size (Fig. 1a). The plots for $q=0$ reveal that when sample size is greater than 100, the bacterial biota in Subject D is more diverse than that in Subject C, and the difference is significant when the sample size exceeds 400. By contrast, for measures of $q=1$ and 2, the direction is reversed. This pattern implies that the proportion of rare branch segments (those branches with low node abundances) for the bacterial biota in Subject D is higher than Subject C, whereas for Subject C the proportion of common/dominant branch segments (those branches with high node abundances) is higher. For each subject the sampling curve for Faith's PD increases steeply with sample size, but the curves for $q=1$ and 2 level off beyond the reference sample, illustrating that higher-order ^{q}PD measures are increasingly dominated by the frequencies of the more common branches, and are therefore less sensitive to sampling effects. The curve for $q=2$ tends to stabilize very quickly with narrow confidence intervals. This is because the phylogenetic measure of $q=2$ is mainly determined by the long branches with very high node abundances, and all these species/lineages can be observed in samples with relatively small sizes

Step 2: *Construct a sample completeness curve to link sample-size- and coverage-based rarefaction/extrapolation curves* (*Fig. 1b*).—The sample completeness curve depicts how sample completeness (measured by sample coverage) increases with sample size with 95% confidence intervals for each habitat up to the maximum size of 500. The plots show that except for initial sample sizes, the coverage for the bacterial biota of Subject C is higher than that of Subject D. The sample coverage curves provide a bridge between sample-size- and coverage-based sampling curves; see Step 3.

Step 3: *Compare coverage-based rarefaction/extrapolation curves up to a maximum coverage* (*Fig. 1c*)—From the sample completeness curve (Fig. 1b), when sample size is increased from the reference sample to 500, the sample coverage for Subject C is increased from 91.0% to 97.2%; for Subject D, it is increased from 81.2% to 92.8%. Within these ranges of sample completeness, the coverage-based sampling curves with 95% confidence intervals for $q=0$, 1, and 2 (Fig. 1c) compares two equally complete samples. The comparison between the two bacterial assemblages generally exhibits a pattern consistent with that found from sample-size-based curves (Fig. 1a),

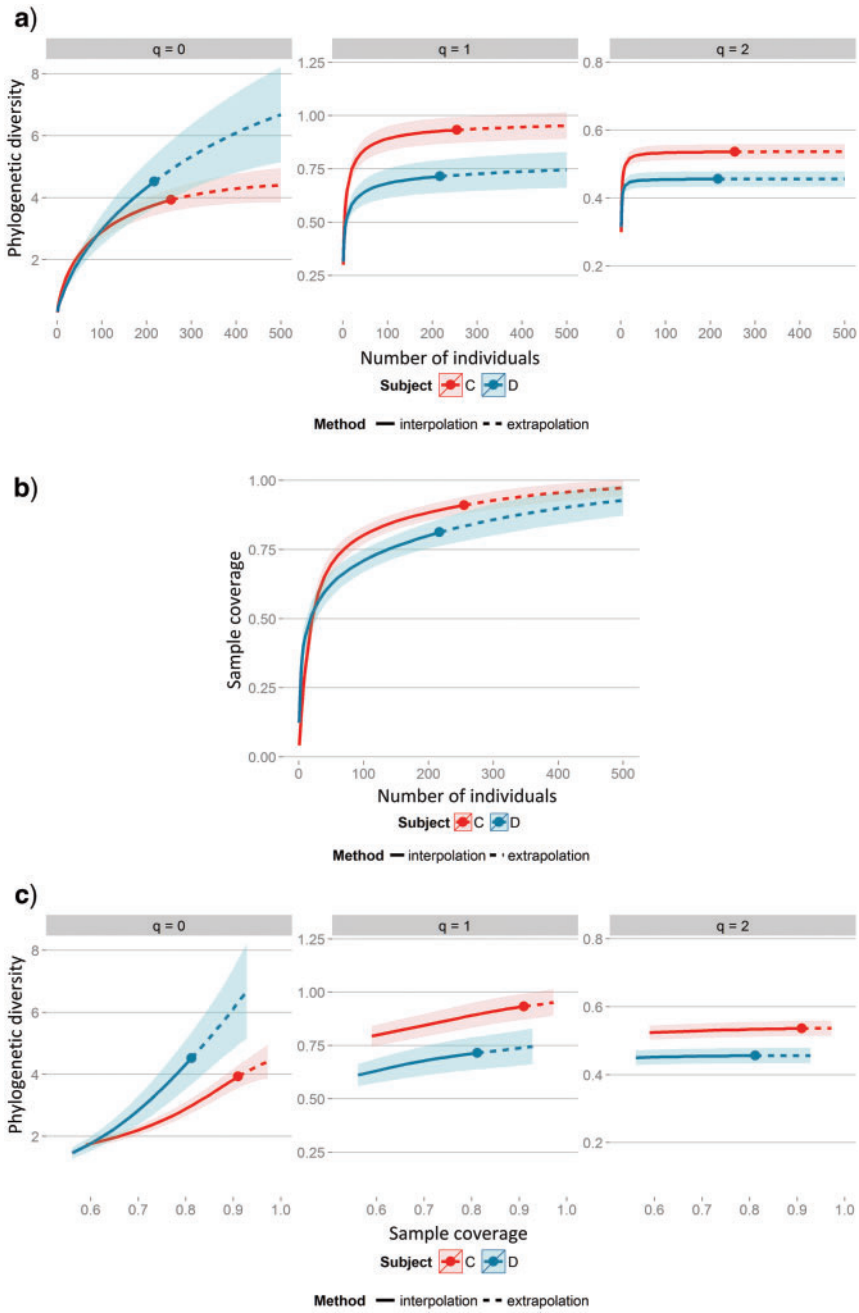

FIGURE 1. Three types of sampling curves for comparing phylogenetic diversity based on the bacterial data from the human distal esophagus of two subjects (Pei et al. 2004). Reference samples are denoted by solid dots. The 95% confidence bands (shaded areas) are obtained by a bootstrap method based on 200 replications. a) Sample-size-based rarefaction (solid lines) and extrapolation (dotted lines) of the phylogenetic measure ^{q}PD for order $q=0$ (left panel), $q=1$ (middle panel), and $q=2$ (right panel), up to the maximum sample size of 500 sequences. b) The sample completeness curve depicts the sample coverage for rarefied samples (solid line) and extrapolated samples (dashed line) as a function of sample size. Each of the curves is extrapolated up to the maximum sample size of 500. c) Coverage-based rarefaction (solid lines) and extrapolation (dotted lines) for the phylogenetic measure ^{q}PD for order $q=0$ (left panel), $q=1$ (middle panel), and $q=2$ (right panel). The curves are plotted up to the coverage 97.2% for Subject C, and to 92.8% for Subject D; these maximum coverage values are obtained from the sample completeness curve for a sample size of 500 for each subject.

but the comparison basis is changed from standardized sample size to sample completeness of the two assemblages. For any standardized fraction of each assemblage, the phylogenetic diversity curve for Subject D lies above the curve of Subject C for the measure with $q=0$; but for the measure with $q=1$ and 2, the ordering is reversed. The two curves with $q=2$ become essentially

flat. An advantage of using the coverage-based curves is that the confidence intervals of the two subjects in Fig. 1c do not overlap, implying a significant difference in diversity of all orders for any given standardized fraction of assemblages. This significant difference between the two microbial biotas for both $q=1$ and $q=2$ is valid not only for the coverage plotted in Fig. 1c but also for any range of coverage up to complete coverage. The validity of this conclusion is also supported by examining the nonoverlapped confidence intervals for the estimated asymptotes of the diversities with $q=1$ and $q=2$ in Table 2. For $q=0$, the significant difference holds up to a coverage of 92.8% (Fig. 1c). However, the two asymptotic estimates (for complete coverage) for $q=0$ cannot be compared due to the limitation of the extrapolated range for measures of low diversity orders; see *Discussion*.

Discussion

In this article, we have developed a general rarefaction and extrapolation framework and estimation method for a class of abundance-based phylogenetic diversity measures (developed by Chao et al. 2010) from a given fixed reference point that is ancestral to all taxa considered in the study. Since our development represents an extension of the rarefaction and extrapolation of Hill numbers to take phylogeny into account, we now have established a unified framework based on rarefaction/extrapolation for assessing and comparing species/taxonomic diversity and phylogenetic diversity across multiple assemblages. In summary, for the analysis of sampling biodiversity data, we suggest presenting two types (sample-size- and coverage-based) of rarefaction and extrapolation for both ordinary Hill numbers and the phylogenetic diversity for the three measures of $q=0$, 1, and 2. As an example, we show the corresponding two types of sampling curves based on Hill numbers (Online Appendix F, available on Dryad) for the two bacterial assemblages and compared them with the curves based on the phylogenetic diversity. In our analysis, we also provide estimated asymptotes for measures of $q=0$, 1, and 2 (Table 2); see below for discussion for the inferences of the estimated asymptotes.

For rarefaction, our proposed estimators work well for any order $q\geqslant 0$ (Table 1). For extrapolation, the prediction bias depends on the extrapolated range and order q. When $q\geqslant 1$, the extrapolated estimator is nearly unbiased for all extrapolated sample sizes, so the extrapolation can be reliably extended to the asymptote if data are sufficient. However, for $q<1$, the extrapolated estimators typically underestimate, and the magnitude of the prediction bias generally increases with the prediction range; the extrapolation is reliable only up to no more than twice the reference sample size. This finding is generally consistent with that for rarefaction and extrapolation for Hill numbers (Chao et al. 2014b) and for Faith's PD (Chao et al. 2015).

As noted above, the extrapolation of the phylogenetic diversity of orders $q=1$ and 2, but not necessarily Faith's PD ($q=0$), can often be reliably extended to infinity or complete coverage to reach the estimated asymptote. Consequently, comparisons for diversity of $q\geqslant 1$ can be made for any sample size up to infinity and for any degree of coverage up to complete coverage if data are sufficient. For example, the estimated asymptotes with 95% confidence intervals (Table 2) can be compared between the two subjects for the measure with $q=1$ and with $q=2$. This is because when the diversity order q is greater than or equal to 1, rare species/lineages (those with low abundances) have less relative impact on these diversities, thus we generally can infer these diversities up to asymptotes and compare them across communities. On the other hand, like the inferences for species richness in hyper-diverse assemblages, sampling data often do not provide sufficient information to accurately infer the sum of undetected branch length ($q=0$ measure) unless strong assumptions or parametric models are made. Our estimated total branch length (Table 2) theoretically is a lower bound (Chao et al. 2015) and thus cannot be compared across assemblages. In this case, fair comparison of Faith's PD across multiple assemblages can be made by standardizing sample completeness (i.e., comparing Faith's PD for a standardized fraction of population) based on coverage-based rarefaction and extrapolation sampling curves up to a maximum degree of fraction of population; beyond that, the data do not contain sufficient information to supply accurate estimates for fair comparison.

In Fig. 1a, the rarefaction of Faith's PD shows that the curves intersect, implying that the ranking of samples differs and depends on the sample size. For species diversity, Lande et al. (2000) showed that intersection occurs when the community with lower actual species richness has larger Gini–Simpson index, which represents the initial slope of a species accumulation curve. Chao et al. (2015) extended a species accumulation curve to a PD accumulation curve. The initial slope of a PD accumulation curve can be readily proved to be Rao's quadratic entropy (Online Appendix B, available on Dryad), which is a generalization of the Gini–Simpson index. Therefore, we can extend the arguments of Lande et al. (2000) to a phylogenetic version: two PD accumulation curves intersect when the community with lower true PD has larger quadratic entropy.

All the proposed analytic estimators for rarefied/extrapolated samples as well as asymptotic estimators are computed from a given fixed reference point that is ancestral to all taxa considered in the study. If the phylogenetic diversity is going to have an objective interpretation that can be compared across studies, it has to have a meaningful reference point that depends on the question being asked, not on the sampled species. Because different investigation may require different reference points, we suggest reporting estimates as a

function of the time or age of the reference point. This is a recommendation to use a diversity profile as a function of "time parameter" (or temporal perspective) made by Chao et al. (2010). In our rarefaction/extrapolation analyses of bacterial sequence data and bat species data (Online Appendix E, available on Dryad), without loss of generality, all our reference points were chosen as the age of the root of all observed taxa. Although the root of the observed taxa varies with sampling data, we can easily transform all our estimates to those for any more basal reference point (Online Appendix G, available on Dryad), facilitating comparisons of diversities for various reference points.

In this article, our derivation is focused on the sampling data obtained when a sample of individuals is randomly selected from an assemblage. In some biological surveys, sampling units, instead of individuals, are randomly selected. The sampling unit is often a trap, net, quadrat, plot, or timed survey. In this sampling scheme, the abundance of each species is often not recorded; only its incidence (presence/absence or detection/nondetection) in each sampling unit is recorded. Although our article deals primarily with abundance data, parallel derivations can be extended to formulate models and derive rarefaction/extrapolation of the corresponding incidence-data-based phylogenetic diversity (Online Appendix H, available on Dryad).

All the estimators in the rarefaction and extrapolation of Hill numbers can be computed from the online freeware application iNEXT (iNterpolation/EXTrapolation), and all corresponding phylogenetic versions are featured in iNextPD (iNterpolation/EXTrapolation for phylogenetic diversity). Both iNEXT and iNextPD can be downloaded from http://chao.stat.nthu.edu.tw/wordpress/software_download/.

We have proposed in this article the use of rarefaction and extrapolation to fairly compare within-assemblage phylogenetic diversity across assemblages. The between-assemblage information is not used and thus phylogenetic differentiation among assemblages is not addressed. Recently, Chiu et al. (2014) developed diversity decomposition of abundance-based phylogenetic gamma diversity measures into alpha and beta (within- and between-group) components. The rarefaction and extrapolation for phylogenetic beta diversity and the associated similarity and differentiation measures merit further research.

Supplementary Material

Data available from the Dryad Digital Repository: http://dx.doi.org/10.5061/dryad.qk58h.

Funding

This work was supported by the Taiwan Ministry of Science and Technology under Contracts 103-2628-M007-007 and 104-2628-M007-003.

Acknowledgments

The authors thank the Editor-in-Chief (Frank Anderson), the Associate Editor (Dan Faith), Lou Jost, David Nipperess, and Erick Matsen for carefully reading an earlier version of the paper and providing very helpful and thoughtful comments and suggestions. Part of the material in this paper is based on the Ph.D. work of the first author under the supervision of the second author. T.C.H. is supported by a post-doctoral fellowship, Taiwan Ministry of Science and Technology.

References

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

Cadotte M.W., Davies T.J., Regetz J., Kembel S.W., Cleland E., Oakley T.H. 2010. Phylogenetic diversity metrics for ecological communities: integrating species richness, abundance and evolutionary history. Ecol. Lett. 13:96–105.

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

Chao A., Chiu C.H., Hsieh T.C., Davis T., Nipperess D.A., Faith D.P. 2015. Rarefaction and extrapolation of phylogenetic diversity. Method Ecol. Evol. 6:380–388.

Chao A., Chiu C.H., Jost L. 2010. Phylogenetic diversity measures based on Hill numbers. Phil. Trans. Roy. Soc. B 365:3599–3609.

Chao A., Chiu C.H., Jost L. 2014a. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. Annu. Rev. Ecol. Evol. Syst. 45:297–324.

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

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

Chao A., Wang Y.T., Jost L. 2013. Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. Method Ecol. Evol. 4:1091–1100.

Chiu C.H., Jost L., Chao A. 2014. Phylogenetic beta diversity, similarity, and differentiation measures based on Hill numbers. Ecol. Monogr. 84:21–44.

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. J. Plant Ecol. 5:3–21.

DeSantis T.Z., Hugenholtz P., Larsen N., Rojas M., Brodie E.L., Keller K., Huber T., Dalevi D., Hu P., Andersen G.L. 2006. Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. Appl. Environ. Microb. 72:5069–5072.

Faith D.P. 1992. Conservation evaluation and phylogenetic diversity. Biol. Conserv. 61:1–10.

Faith D.P. 2013. Biodiversity and evolutionary history: useful extensions of the PD phylogenetic diversity assessment framework. Ann. NY Acad. Sci. 1289:69–89.

Faith D.P., Richards Z.T. 2012. Climate change impacts on the tree of life: changes in phylogenetic diversity illustrated for Acropora corals. Biology 1:906–932.

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. J. Stat. Comput. Sim. 66:101–111.

Gotelli N.J., Colwell R.K. 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. Ecol. Lett. 4:379–391.

Helmus M.R., Keller W., Paterson M.J., Yan N.D., Cannon C.H., Rusak J.A. 2010. Communities contain closely related species during ecosystem disturbance. Ecol. Lett. 13(2):162–174.

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

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

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

Kembel S.W., Eisen J.A., Pollard K.S., Green J.L. 2011. The phylogenetic diversity of metagenomes. PLoS One 6(8):e23214.

Kembel S.W., Jones E., Kline J., Northcutt D., Stenson J., Womack A.M., Bohannan B.J., Brown G.Z., Green J.L. 2012. Architectural design influences the diversity and structure of the built environment microbiome. ISME J. 6:1469–1479.

Lande R., DeVries P., Walla T. 2000. When species accumulation curves intersect: implications for ranking diversity using small samples. Oikos 89:601–605.

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. Appl. Environ. Microb. 75(15):5111.

MacArthur R.H. 1965. Patterns of species diversity. Biol. Rev. 40: 510–533.

McCoy C.O., Matsen F.A. 2013. Abundance-weighted phylogenetic diversity measures distinguish microbial community states and are robust to sampling depth. PeerJ 1:e157.

McPeek M.A., Miller T.E. 1996. Evolutionary biology and community ecology. Ecology 77:1319.

Nipperess D.A., Matsen F.A. 2013. The mean and variance of phylogenetic diversity under rarefaction. Method Ecol. Evol. 4: 566–572.

Pei Z., Bini E.J., Yang L., Zhou M., Francois F., Blaser M.J. 2004. Bacterial biota in the human distal esophagus. Proc. Natl. Acad. Sci. USA. 101:4250–4255.

Rao C.R. 1982. Diversity and dissimilarity coefficients: a unified approach. Theor. Popul. Biol. 21:24–43.

Ricotta C., Bacaro G., Marignani M., Godefroid S., Mazzoleni S. 2012. Computing diversity from dated phylogenies and taxonomic hierarchies: does it make a difference to the conclusions? Oecologia 170:501–506.

Schloss P.D., Westcott S.L., Ryabin T., Hall J.R., Hartmann M., Hollister E.B., Lesniewski R.A., Oakley B.B., Parks D.H., Robinson C.J. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. Appl. Environ. Microb. 75: 7537–7541.

Warwick R.M., Clarke K.R. 1995. New 'biodiversity' measures reveal a decrease in taxonomic distinctness with increasing stress. Mar. Ecol. Prog. Ser. 129:301–305.

Webb C.O., Gilbert G.S., Donoghue M.J. 2006. Phylodiversity-dependent seedling mortality, size structure, and disease in a Bornean rain forest. Ecology 87:S123–S131.