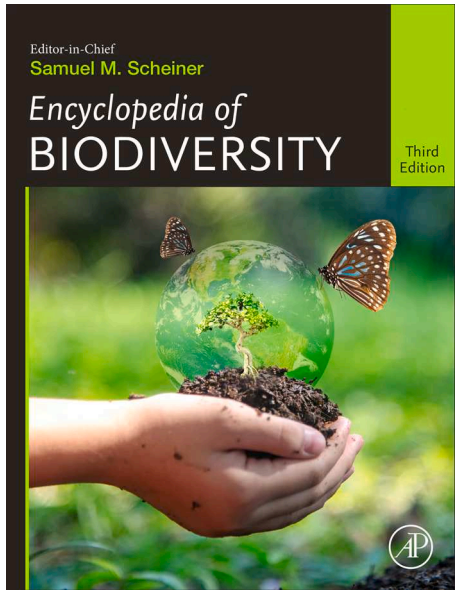

Gotelli, Nicholas J., Chao, Anne and Colwell, Robert K. (2024) Measuring and Estimating Species Richness, Taxonomic and Phylogenetic Diversity, and Related Biotic (Dis)similarity Indices From Sampling Data. In: Scheiner Samuel M. (eds.) *Encyclopedia of Biodiversity 3rd edition,* vol. 5, pp. 314–339. Oxford: Elsevier.

http://dx.doi.org/10.1016/B978-0-12-822562-2.00104-3

Measuring and Estimating Species Richness, Taxonomic and Phylogenetic Diversity, and Related Biotic (Dis)similarity Indices From Sampling Data

Nicholas J Gotelli, Department of Biology, University of Vermont, Burlington, VT, United States
Anne Chao, Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan
Robert K Colwell, Department of Ecology and Evolutionary Biology, University of Connecticut, Storrs, CT, United States; University of Colorado Museum of Natural History, Boulder, CO, United States; and Department of Ecology, Federal University of Goiás, Goiânia, GO, Brazil

This is an update of Nicholas J. Gotelli, Anne Chao, Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data, Editor(s): Simon A Levin, Encyclopedia of Biodiversity (Second Edition), Academic Press, 2013, Pages 195–211, ISBN 9780123847201, https://doi.org/10.1016/B978-0-12-384719-5.00424-X.

Abstract

This article reviews modern statistical approaches to estimating species/taxonomic diversity (including species richness), phylogenetic diversity, and related biotic (dis)similarity measures from empirical samples of species abundance or incidence. Key methods include asymptotic diversity estimation, non-asymptotic standardization via rarefaction and extrapolation, diversity decomposition (alpha, beta, and gamma), and related (dis)similarity measures. Hill numbers are presented as unifying measures of species diversity that take into account species richness and species relative abundance. Under the framework of Hill numbers, all key methods can be extended to account for phylogenetic and functional diversity of biological assemblages. These statistical methods control for several kinds of sampling effects in biodiversity data and can be applied to many important questions in ecology.

Key Points

- Hill numbers, which integrate species richness and species relative abundance into a class of measures (parameterized by a diversity order $q \geq 0$), represent a unified framework for quantifying species/taxonomic diversity.
- The framework of Hill numbers can be generalized to phylogenetic and functional diversity. (This article focuses primarily on the phylogenetic generalization.)
- Hill numbers also provide a unified approach to decomposing diversity (alpha, beta, and gamma) and constructing normalized taxonomic and phylogenetic (dis)similarity measures.
- For undersampled data, species richness (Hill number of order $q = 0$) is a difficult parameter to accurately estimate. Nonparametric estimators, which use information on the observed frequencies of rare species, are recommended for estimating the minimum number of species.
- Abundance-sensitive diversity and (dis)similarity measures of orders $q \geq 1$ and their phylogenetic versions are less sensitive to undersampling. Nearly unbiased estimators exist for measures of $q = 2$ (dominated by very abundant species). Undersampling bias for measures of $q = 1$ (focusing on abundant/common species) can be largely removed by asymptotic statistical estimation.
- When data are not sufficient to provide accurate asymptotic estimators (typically for measures of $q = 0$, species richness), non-asymptotic standardization via interpolation and extrapolation is recommended for valid comparisons of diversity and (dis)similarity measures across multiple datasets.

Glossary

Assemblage A collection of species populations—often taxonomically, phylogenetically, or functionally related—that is sampled within a specified spatial and temporal domain.

Biotic (dis)similarity A measure of the degree to which two or more samples or assemblages are (dis)similar in species composition. Familiar biotic (dis)similarity indices include Sørensen's, Jaccard's, Horn's, and Morisita's indices. These measures can be extended to their phylogenetic versions.

Evenness The extent of equity among species abundances.

Extrapolation A statistical prediction method of extrapolating a reference sample in order to standardize the comparison of diversity on the basis of a common sample size or equivalent sample coverage.

Hill numbers A family of diversity measures developed by Mark Hill. Hill numbers quantify diversity in units of equivalent numbers of equally abundant species.

Individual-based (abundance) data A common form of data in biodiversity surveys. The data set consists of a vector of the abundances of different species. This data structure is used when an investigator randomly samples individual organisms in a biodiversity survey.

iNEXT Abbreviation for interpolation and extrapolation, a non-asymptotic standardization method based on interpolation (rarefaction) and extrapolation.

Nonparametric asymptotic species-richness estimators Estimators of total species richness (including Chao1, Chao2, abundance-based coverage estimator (ACE), incidence-based coverage estimator (ICE), and the jackknife), which do not assume a particular form of the species abundance distribution (such as a log-series, log-normal, or uniform distribution). Instead, these methods use information on the frequency of rare species in a sample to estimate the number of undetected species in the assemblage.

Phylogenetic diversity Adjusted diversity measures that take into account the degree of relatedness determined by Linnean taxonomic classification or evolutionary history among a set of species in an assemblage. Other things being equal, an assemblage of closely-related species is less phylogenetically diverse than a set of distantly-related species.

Rarefaction A statistical interpolation method of rarefying or thinning a reference sample in order to standardize the comparison of biological diversity on the basis of a common sample size or equivalent sample coverage.

Reference sample Observed data in the form of a list of species abundance/frequency records in a single sampling unit (for abundance data) or species detection/non-detection records in several sampling units (for incidence data).

Sample-based (incidence) data A common form of data in biodiversity surveys. The data set consists of a set of sampling units (such as plots, quadrats, traps, or transect lines). The incidence or presence of each species (but not its abundance) is recorded for each sampling unit. Individual-based (abundance) data collected from sampling units can be converted to incidence data, but not the reverse.

Sample coverage The fraction of an assemblage's individuals (for abundance data) or incidences (for incidence data) that belong to the species that were observed in the sample. Sample coverage is an objective measure of sample completeness developed by Alan Turing in his cryptographic work during WWII.

Species abundance distribution (SAD) A complete distribution of the abundance (number of individuals) of every species in an assemblage. A general pattern of SAD is that most species are rare, and only a few species are common. Thus many uncommon species may be missing from a limited sample of individuals from an SAD.

Species accumulation curve A curve of rising biodiversity in which the x-axis is the number of sampling units (individuals or samples) for an assemblage and the y-axis is the observed species richness. The species accumulation curve rises monotonically to an asymptotic maximum number of species, for a given assemblage or habitat. Beyond this asymptote, additional sampling units do not yield additional species.

Species/taxonomic diversity A measure of diversity that incorporates both the number of species in an assemblage and evenness of their relative abundances. Many species diversity indices can be converted by an algebraic transformation to Hill numbers.

Species richness The total number of species in an assemblage or a sample. Species richness in an assemblage is difficult to estimate reliably from sample data because it is very sensitive to the number of individuals and the number of sampling units collected. In the Hill number framework, species richness is a diversity of order 0 (which means it disregards species abundances).

Undersampling A downward sampling bias in species richness counts arising from insufficient sampling effort to detect all species in an assemblage or habitat.

Introduction

Measuring Biological Diversity

The notion of biological diversity is pervasive at levels of organization ranging from the expression of heat-shock proteins in a single fruit fly to the production of ecosystem services by a terrestrial ecosystem threatened by climate change. How can one quantify diversity in meaningful units across such different levels of organization? This article describes a basic statistical framework for quantifying diversity and making meaningful inferences from samples of diversity data.

In very general terms, a collection of "elements" is considered, each of which can be uniquely assigned to one of several distinct "types" or categories. In community ecology, the elements typically represent the individual organisms, and the types represent distinct species. These definitions are generic, however, and typically are modified for different kinds of diversity studies. For example, paleontologists often cannot identify fossils to the species level, so they might study diversity at higher taxonomic levels, such as genera or families. Population geneticists and molecular biologists might be interested in more fine-scale "omics": classifications of biological materials on the basis of unique DNA sequences or operational taxonomic units (OTU, genomics), expressed mRNA molecules (transcriptomics), proteins (proteomics), or metabolic products (metabolomics). Ecosystem ecologists might be concerned not with individual molecules, genotypes, or species, but with broad functional groups (producers, predators, and decomposers) or specialized ecological or evolutionary life forms (trees, understory forest herbs, or filter-feeding mollusks). Archeologists might focus on flaking styles of projectile points. However, to keep things simple, this article will refer throughout to "species" as the distinct categories of biological classification.

Although the sampling unit is often thought of as the individual organism, many species, such as clonal plants or colonial invertebrates, do not occur as distinct individuals that can be counted. In other cases, the individual organisms, such as aquatic invertebrate larvae, marine phytoplankton, or soil microbes are so abundant that they cannot be practically counted. In these cases, the elements of biodiversity will correspond not to individual organisms, but to the sampling units (traps, quadrats, and sighting records) that ecologists use to record the presence or presumed absence of a species.

Species Richness and Traditional Species Diversity Metrics

The number of species in an assemblage is the most basic and natural measure of diversity. Many important theories in community ecology, including island biogeography, intermediate disturbance, keystone and foundational species effects, neutral theory, and metacommunity dynamics make quantitative predictions about species number that can be tested with field observations and experiments in community ecology. From the applied perspective, species richness is sometimes considered the ultimate "score card" in efforts to preserve biodiversity in the face of increasing environmental pressures and climate change resulting from human activity. Species losses can occur from extinction, and increases in the number of species can reflect deliberate and accidental introductions or range shifts of species driven by climate change.

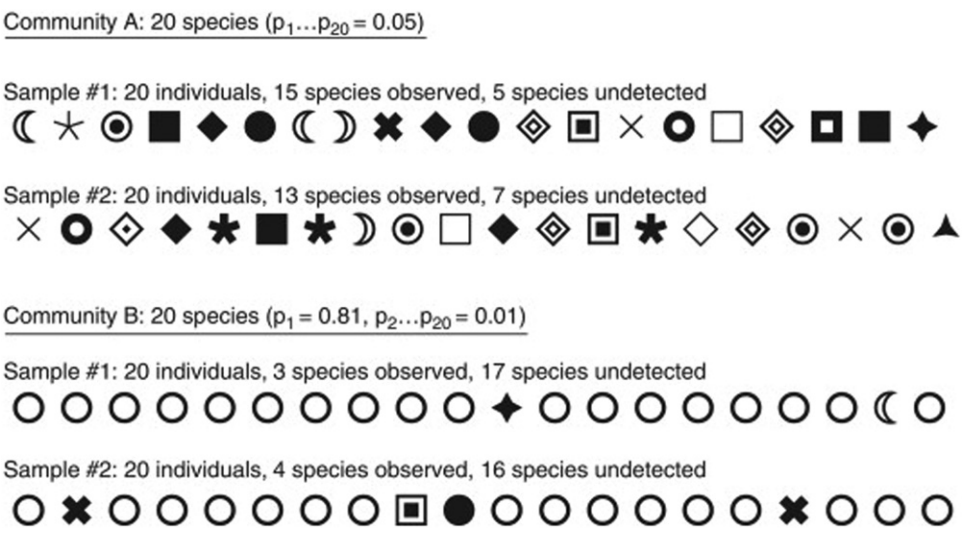

Fig. 1 Species richness sampling in a hypothetical walk through the woods. Each different symbol represents one of 20 distinct species, and each row contains 20 characters, representing the first 20 individual trees that might be encountered in a random sample. Community A is maximally even, with each of the 20 species comprising 5% of the total abundance. In this assemblage, the two samples of 20 individual trees yielded 15 and 13 species, respectively. Community B is highly uneven, with one species (the open circle) representing 81% of the total abundance, and the remaining 19 species contributing only 1% each. In this assemblage, the two samples of 20 individual trees yielded only three and four species, respectively.

Although species richness is a key metric, it is not the only component of species diversity. Consider two woodlands, each with 20 species of trees. In the first woodland, the 20 species are equally abundant, and each species comprises 5% of the total abundance. In the second woodland, one dominant species comprises 81% of the total abundance, and each of the remaining 19 species contributes only 1% to the total. Although both woodlands contain 20 species, a visitor to the first woodland would encounter most of the different tree species in a brief visit, whereas a visitor to the second woodland might encounter mostly the single dominant species (Fig. 1).

Thus, a comprehensive measure of species diversity should include components of both species richness and the relative abundances of the species that are present. Such measures are referred to in this article as "traditional" species/taxonomic diversity measures. Ecologists have used dozens of different traditional diversity measures, all of which assume (1) individuals within a species are equivalent, (2) all species are "equally different" from one another and thus receive equal weighting, and (3) diversity is measured in appropriate units (individuals, biomass, and percentage cover are commonly used).

Phylogenetic and Functional Diversity

As noted in the previous section, the first assumption of traditional diversity metrics (individuals within a species are equivalent) can be relaxed by changing the operational definition of "species" to other categories of interest. The second assumption (all species are equally different from one another) ignores aspects of phylogenetic or functional diversity but can also be incorporated through a generalization of traditional diversity metrics.

For example, consider two woodlands with identical tree species richness and relative abundance but with no shared species. The species in the first woodland are all closely related oaks in the same genus (*Quercus*). The species in the second woodland are a diverse mix of oaks (*Quercus*) and maples (*Acer*), as well as more distantly related pines (*Pinus*). Traditional species diversity metrics would be identical for both woodlands, but it is intuitive that the second woodland is more diverse (see Fig. 2 for another example).

The traditional concept of diversity can therefore be extended to consider differences among species. All else being equal, an assemblage of phylogenetically distant or functionally dissimilar species is more diverse than an assemblage of closely related or functionally similar species. Differences among species can be based directly on their evolutionary histories, either in the form of a Linnean taxonomic hierarchy, on a phylogenetic tree (phylogenetic diversity (PD)) or based on their function (functional diversity (FD)). Because they weight each species by a measure of its Linnean taxonomic classification, phylogeny, or function, these metrics relax the second assumption discussed above in the section Species Richness and Traditional Species Diversity Metrics: all species are equally different from one another.

Biotic (Dis)similarity

These concepts of species diversity apply to metrics that are used to quantify the diversity of single assemblages. However, the concept of diversity can also be applied to the comparison of multiple assemblages. Suppose again that a person visits two woodlands, each of which has 10 trees species, each species contributing 10% to the abundance of individual trees within the

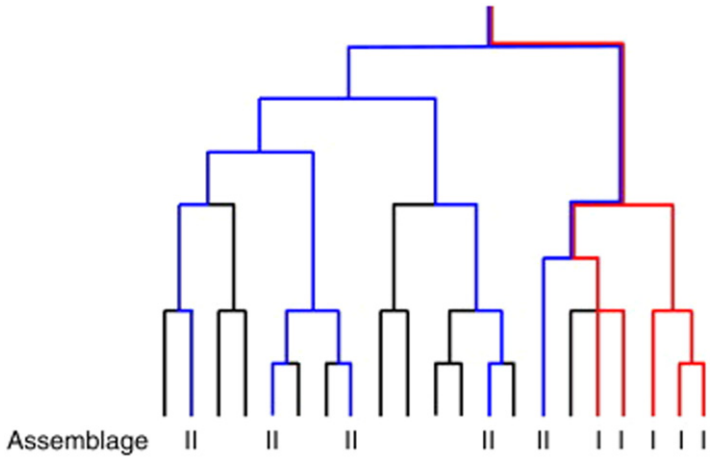

Fig. 2 Phylogenetic diversity in species composition. The branching diagram is a hypothetical phylogenetic tree. The ancestor of the entire assemblage is the "root" at the top, with time progressing toward the branch tips at the bottom. Each node (branching point) represents a speciation or divergence event, and the 21 branch tips illustrate the 21 extant species. Extinct species or lineages are not illustrated. The five species in Assemblage I (with spanning tree in red branches) represent an assemblage of five closely related species (they all share a quite recent common ancestor). The five species in Assemblage II (with spanning tree in blue branches) represent an assemblage of five distantly related species (they all share a much older common ancestor). All other things being equal, the community of distantly related species would be considered more phylogenetically diverse than the community of closely related species.

woodland. Thus, in terms of species richness and species diversity, the two woodlands are identical. However, the two woodlands may differ in their species composition. At one extreme, they may have no species in common, so they are biologically distinct, in spite of having equal species richness and species diversity. At the other extreme, if the list of tree species in the two woodlands and their relative abundance or incidence are the same, they are identical in all aspects of diversity (including taxonomic, phylogenetic, and functional diversity). More typically, the two woodlands might have a certain number of species found in both woodlands and a certain number that are found in only one.

Biotic similarity quantifies the extent to which two or more sites are similar in their species composition and relative abundance distribution. The concept of biotic similarity is important at large spatial scales for the designation of biogeographic provinces that harbor distinctive species assemblages with both endemic and shared elements. Biotic similarity is also a key concept underlying the measurement of beta diversity, the turnover in species composition among a set of sites. In an applied context, biotic similarity indices can quantify the extent to which distinct biotas in different regions have become homogenized through loss of endemic species and the introduction and spread of nonnative species. Differences among species in evolutionary histories and functional trait values can also be incorporated in similarity measures.

Bias in the Estimation of Diversity

The true species richness and relative abundances in an assemblage are unknown in most applications. Thus, species richness, species diversity, biotic similarity, and their phylogenetic versions must be estimated from samples taken from the assemblage. When sample size is not sufficiently large to observe all species, the unobserved species are undersampled, and—as a consequence—the relative abundance of observed species, on average, is overestimated.

Because biotic diversity at all levels of organization is often high and biodiversity sampling is labor intensive, these biases are usually substantial. Even the simplest comparison of species richness between two samples is complicated unless the number of individuals is identical in the two samples (which it never is) or the two samples represent the same degree of coverage (completeness) in sampling. Ignoring sampling effects may obscure the influence of overall abundance or sampling intensity on species richness. Attempts to adjust for sampling differences by algebraic rescaling (such as dividing S by n or by sampling effort) lead to serious distortions and gross overestimates of species richness for small samples. Thus, an important general objective in diversity analysis is to reduce undersampling bias and to adjust for the effect of undersampled species on the estimation of diversity and similarity measures. Because sampling variation is an inevitable component of biodiversity studies, it is equally important to assess the variance (or standard error) of an estimator and provide a confidence interval that reflects sampling uncertainty.

The Organization of Biodiversity Sampling Data

This article introduces a common notation for describing biodiversity data (Colwell *et al.*, 2012). Consider an assemblage consisting of z_+ total individuals, each belonging to one of S distinct species. Species i has z_i individuals, so that $\sum_{i=1}^{S} z_i = z_+$. The

relative abundance or frequency p_i of species i is $p_i = z_i/z_+$, so that $\sum_{i=1}^{S} p_i = 1$. Note here that z_+, S, z_i, and p_i represent the "true" underlying total abundance, species richness, absolute abundance, and relative abundance of species. These quantities are unknowns, but they can be estimated, and one can make statistical inferences by taking random samples of data from such an assemblage. This article distinguishes between two sampling structures.

Individual-Based (Abundance) Data

The *reference sample* is a collection of n individuals, drawn at random from the assemblage with z_+ total individuals (Colwell *et al.*, 2012). In the reference sample, a total of S_{obs} species is observed, with X_i individuals observed for species i, so that $\sum_{i=1}^{S} X_i = n$ (only species with $X_i > 0$ contribute to the sum). Thus, the data consist of a single vector of length S, whose elements are the observed abundances of the individual species X_i. In this vector, there are S_{obs} nonzero elements.

The *abundance frequency count* f_k is defined as the number of species each represented by exactly k individuals in the reference sample. Thus, f_1 is the number of species represented by exactly one individual ("singletons") in the reference sample, and f_2 is the number of species represented by exactly two individuals ("doubletons"). In this terminology, f_0 is the number of *undetected species*: species that are present in the assemblage of z_+ individuals and S species but were represented by 0 individuals (and thus not detected) in the reference sample of n individuals and S_{obs} species. Therefore, $S_{obs} + f_0 = S$.

Sample-Based (Incidence) Data

The reference sample for incidence data consists of a set of R replicated *sampling units* (traps, plots, quadrats, search routes, etc.). In a typical study, these sampling units are deployed randomly in space within the area encompassing the assemblage. However, in a temporal study of diversity, the R sampling units would be deployed in one place at different independent points in time (such as an annual breeding bird census at a single site). Within each sampling unit, the detection or non-detection of each species is recorded, but abundances or counts of the species present are not used. The data are organized as a species-by-sampling-unit *incidence matrix*, in which there are $i = 1$ to S rows (species), $j = 1$ to R columns (sampling units), and the matrix entry $W_{ij} = 1$ if species i is detected in sampling unit j, and $W_{ij} = 0$ otherwise. If sampling is replicated in both time and space, the data would be organized as a three-dimensional matrix (species $\times$ sites $\times$ times). However, most biodiversity data sets are two dimensional, with either spatial or temporal replication, but not both.

A row sum of the incidence matrix $Y_i = \sum_{j=1}^{R} W_{ij}$ denotes the incidence-based frequency of species i for $i = 1$ to S. Y_i is analogous to X_i in the individual-based abundance vector. Species present in the assemblage but not detected in any sampling unit have $Y_i = 0$. The total number of species observed in the reference sample is S_{obs} (only species with $Y_i > 0$ contribute to S_{obs}).

The *incidence frequency count* Q_k is the number of species each represented exactly $Y_i = k$ times in the incidence matrix sample, $0 \le k \le R$. For the incidence matrix $\sum_{k=1}^{R} kQ_k = \sum_{i=1}^{S} Y_i$, and $S_{obs} = \sum_{k=1}^{R} Q_k$. Thus, Q_1 represents the number of "unique" species (each detected in only one sampling unit) and Q_2 represents the number of "duplicate" species (each detected in exactly two sampling units). The zero frequency Q_0 denotes the number of species among the S species in the complete assemblage that were not detected in any of the R sampling units.

Species Richness Estimation

A simple count of the number of species in a sample is usually a biased underestimate of the true number of species, simply because increasing the sampling effort (through counting more individuals, examining more sampling units, or sampling a larger area) inevitably will increase the number of species observed. The effect is best illustrated in a *species accumulation curve*, in which the x-axis is the number of individuals sampled or sampling units examined and the y-axis is the number of species observed (**Fig. 3**). The first individual sampled always yields one species, so the origin of an abundance-based species accumulation curve is the point [1,1]. If the next individual sampled is the same species, the curve stays flat with a slope of zero. If the next individual sampled is a different species, the curve rises to two species, with an initial slope of 1.0. Samples from the real world fall between these two idealized extremes, and the slope of the curve measured at any abundance level is the probability that the next individual sampled represents a previously unsampled species. The curve is steepest in the early part of the collecting, because the common species in the assemblage are detected relatively quickly. The curve continues to rise as more individuals are sampled, but the slope becomes shallower because progressively more sampling is required to detect the next rare species. As long as the sample size or effort is sufficiently large, all of the species will eventually be sampled, and the curve will flatten out at an asymptote that represents the true species richness for the assemblage. For incidence data, an analogous accumulation curve can be drawn in which the x-axis represents the number of sampling units and the y-axis is the number of species recorded.

Interpolating Species Richness With Rarefaction

A single empirical sample of individuals or a pooled set of sampling units represents one point on the species accumulation curve, but the investigator has no way of directly determining where on the curve this point lies. To compare the richness of two different

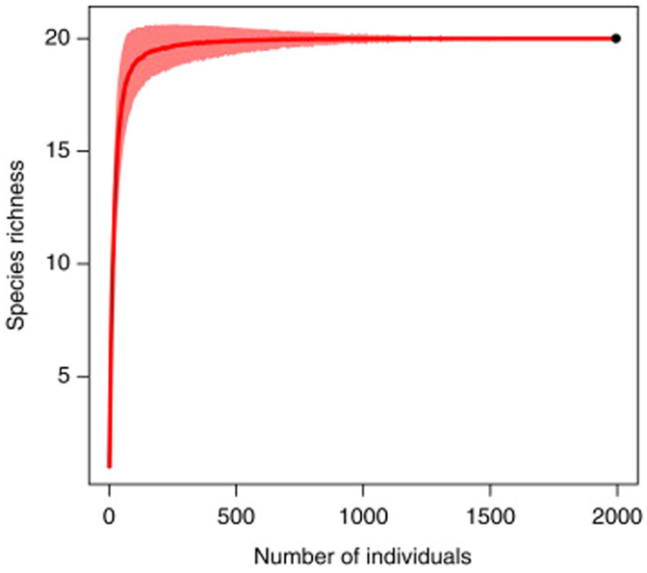

Fig. 3 Species accumulation curve. The curve was generated by assuming an assemblage of 20 species whose relative abundances were created from a broken stick distribution; see Tokeshi (1999) for a description of the broken stick distribution. The *x*-axis is the number of individuals sampled and the *y*-axis is the number of species observed. The species accumulation curve is the smooth red line, which represents the average of 1000 random draws, sampling with replacement, at each level of abundance. The shaded envelope represents a symmetric 95% bootstrap confidence interval, calculated from the estimated variance of the random draws. The shape of this species accumulation curve is typical: it rises rapidly at first as the common species are initially encountered, and then continues to rise very slowly, as much more sampling is needed to encounter all of the rare species. For random samples of 500 or more individuals, it is almost always the case that all 20 species are encountered.

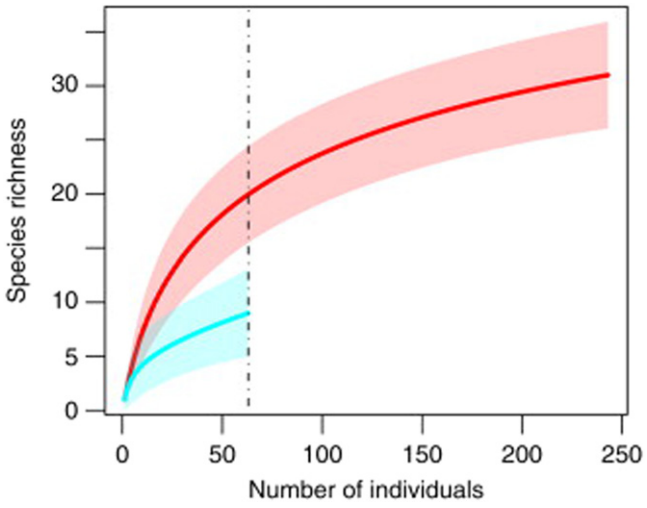

Fig. 4 Standardized comparison of species richness for two individual-based rarefaction curves. The data represent summary counts of carabid beetles that were pitfall-trapped from a set of young pine plantations (< 20 years old; upper curve) and a set of old pine plantations (20–60 years old; lower curve). The solid lines are the rarefaction curves, calculated from Eq. (1), and the shaded polygons are the 95% confidence intervals, calculated from the unconditional asymptotic variance. The young plantation samples contained 243 individuals representing 31 species, and the old plantation samples contained 63 individuals representing nine species. The dashed and dotted vertical line illustrates a species richness comparison standardized to 63 individuals, which was the observed abundance in the smaller of the two data sets. Data from Niemelä J, Haila Y, Halme E, *et al.* (1988). The distribution of carabid beetles in fragments of old coniferous taiga and adjacent managed forest. Annales Zoologici Fennici 25, 107–199.

samples, a conventional approach for abundance data is to standardize to a common number of individuals (Sanders, 1968; Gotelli and Colwell, 2001, 2011). Rarefaction represents an interpolation of a biodiversity sample to a smaller number of individuals for purposes of comparison among samples. Typically, the abundance of the larger sample is rarefied to the total abundance of the smaller sample to determine if species richness (or any other biodiversity index) differs for a common number of

individuals (Fig. 4). For incidence data, rarefaction interpolates between the reference sample and a smaller number of sampling units.

Let $S_{\text{ind}}(m)$ represent the expected number of species in a random sample of m individuals from the reference sample of n individuals ($m < n$). If the true probabilities ($p_1, p_2, \ldots, p_S$) of each of the S species in the assemblage were known, and species frequencies ($X_1, X_2, \ldots, X_S$) follow a multinomial distribution for which the total of all frequencies is n, and cell probabilities ($p_1, p_2, \ldots, p_S$), then

$$S_{\text{ind}}(m) = S - \sum_{i=1}^{S} (1 - p_i)^m.$$

However, the true p_i values are unknown, and there is only the reference sample with observed species abundances X_i. An unbiased estimator for $S_{\text{ind}}(m)$ (Hurlbert, 1971) is

$$\tilde{S}_{\text{ind}}(m) = S_{\text{obs}} - \sum_{X_i > 0} \left[\binom{n - X_i}{m} \Big/ \binom{n}{m} \right]. \quad (1)$$

For incidence-based data, the corresponding equation (Shinozaki, 1963) is:

$$\tilde{S}_{\text{sample}}(r) = S_{\text{obs}} - \sum_{Y_i > 0} \left[\binom{R - Y_i}{r} \Big/ \binom{R}{r} \right],$$

where $r < R$ is the number of sampling units in the rarefied reference sample. Conditional on the observed data, the above two equations imply that the rarefaction formula is obtained by sampling without replacement from the reference sample.

If the area of each sampling unit has been measured, species richness can also be interpolated from a *Coleman curve*, in which the expected species richness on an island (or sample plot) of area a is based on a Poisson model of constant density (individuals per unit area) and is a function of the total area A of the archipelago (or the summed areas of all the sampling units) (Coleman *et al.*, 1982):

$$\tilde{S}_{\text{area}}(a) = S_{\text{obs}} - \sum_{X_i > 0} \left(1 - \frac{a}{A}\right)^{X_i}.$$

Although the variance from the hypergeometric distribution has traditionally been used to calculate a confidence interval for a rarefaction curve, this variance is conditional on the observed sample. It therefore has the undesirable property of converging to zero when the abundance level reaches the reference sample size. More realistically, the observed sample is itself drawn from a much larger assemblage, so the confidence intervals should widen as the reference sample size is reached. This unconditional variance for abundance data can be calculated from an asymptotic formula (Colwell *et al.*, 2012) or is based on a bootstrap method (Chao *et al.*, 2014).

Because all rarefaction curves converge at small sample sizes towards the point [1,1] (for abundance data) or a small number of species (for incidence data), sufficient sampling is necessary for valid comparisons of curves. Although there are no theoretical guidelines, empirical examples suggest that samples of at least 20–50 individuals per sample (and preferably many more) are necessary for meaningful comparisons of abundance-based rarefaction curves.

Rarefaction curves also require comparable sampling methods (forest samples collected from pitfall traps cannot be validly compared to prairie samples collected from baits), well-defined assemblages of discrete, countable individuals (for abundance-based methods), and random, independent sampling of individuals (or sampling units for incidence-based methods). If the sampling of individuals is random and independent, rarefaction provides an unbiased estimate of species richness (and other diversity measures), regardless of any inter-specific or intra-specific spatial structure in the occurrence of individuals. If sampling is not random and independent, rarefaction provides an unbiased estimate of species richness (and other diversity measures) only when there is a random spatial arrangement of individuals.

If the spatial distribution of individuals is intraspecifically clumped in space and individuals are not randomly sampled (e.g., in a quadrat sampling), abundance-based rarefaction will over-estimate species richness, but this problem can be effectively countered by using incidence-based methods (Colwell *et al.*, 2004; Chao and Colwell, 2017). Perhaps the chief disadvantage of rarefaction is that point comparisons force an investigator to rarefy all samples down to the smallest sample size in the data set, so sufficient sampling is important. Otherwise, the magnitude of difference in species richness among assemblages will be much compressed based on a low rarefied sample size (Peet, 1974). However, calculation and comparison of complete rarefaction curves and their extrapolation, with unconditional variances, helps to overcome this problem (Colwell *et al.*, 2012).

Nonparametric Asymptotic Species Richness Estimators

Whereas rarefaction is a method for interpolating species diversity data, asymptotic richness estimators are methods for extrapolating species diversity out to the (presumed) asymptote, beyond which additional sampling will not yield any new species. Three strategies have been used to try to estimate the asymptote of the species accumulation curve (Colwell and Coddington, 1994). *Parametric curve-fitting* uses the shape of the species accumulation curve in its early phase to try and predict the asymptote. Asymptotic functions, such as the negative exponential distribution, the Weibull distribution, the logistic equation, and the Michaelis–Menten equation, are fit (usually with nonlinear regression methods that wrongly assume independence of successive

points along the curve) to the species accumulation data, and the asymptote is then estimated as one of the parameters of this kind of model. The chief problem is that this approach does not work well in comparisons with empirical or simulated data sets, mainly because it does not directly use information on the frequency of common and rare species, but simply tries to forecast the shape of the rising curve. Several different functional forms may fit the same data set equally well, but yield drastically different estimates of the asymptote. Because curve-fitting is not based on a statistical sampling model, the variance of the resulting asymptote cannot be evaluated without further assumptions, and theoretical difficulties arise for model selection.

A second strategy is to use the abundance or incidence frequency counts (f_k or Q_k) and fit them to a *species abundance distribution (SAD)*, such as the log-series or the log-normal distribution. The area under such a fitted curve is an estimate of the total number of species present in the assemblage. The chief weakness of these methods is that simulations show that they work well only when the correct form of the species abundance distribution is already known, but this is never the case for empirical data. Existing statistical models fail to fit empirical data sets well, which often depart from expected values in the frequencies of the rare species. Moreover, there is no guarantee that two different assemblages follow the same kind of distribution, which complicates the comparison of curves.

The most successful methods so far have been *nonparametric estimators* (Colwell and Coddington, 1994), which use the rare frequency counts to estimate the frequency of the missing species (f_0 or Q_0). For incidence data, these estimators are similar to mark-recapture models that are used in demography to estimate the total population size and are based on statistical theorems developed by Alan Turing and I. J. Good from cryptographic analysis of Wehrmacht coding machines during World War II. The basic concept of their theorem is that abundant species—which are certain to be detected in samples—contain almost no information about the undetected species, whereas rare species—which are likely to be either undetected or infrequently detected—contain almost all the information about the undetected species.

If there are many undetectable or "invisible" species in a hyperdiverse assemblage, it will be impossible to obtain a good estimate of species richness. Therefore, an accurate lower bound for species richness is often of more practical use than an imprecise point estimate. Based on the concept that rare species carry the most information about the number of undetected species, the Chao1 estimator uses only the numbers of singletons and doubletons (and the observed richness) to obtain the following lower bound for the expected asymptotic species richness (Chao, 1984):

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

with an associated variance estimator of (if $f_2 > 0$):

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

For incidence data, *Chao2* is the corresponding estimator for species richness. It incorporates a sample-size correction factor $(R-1)/R$ (Chao, 1987):

$$\hat{S}_{\text{Chao2}} = \begin{cases} S_{\text{obs}} + [(R-1)/R]Q_1^2/(2Q_2), & \text{if } Q_2 > 0 \\ S_{obs} + [(R-1)/R]Q_1(Q_1-1)/2, & \text{if } Q_2 = 0 \end{cases}$$

with variance estimator (if $Q_2 > 0$):

$$\hat{var}(\hat{S}_{\text{Chao2}}) = Q_2\left[\frac{A}{2}\left(\frac{Q_1}{Q_2}\right)^2 + A^2\left(\frac{Q_1}{Q_2}\right)^3 + \frac{1}{4}A^2\left(\frac{Q_1}{Q_2}\right)^4\right],$$

where $A = (R-1)/R$. When $f_2 = 0$ in the Chao1 estimator or $Q_2 = 0$ in the Chao2 estimator, the above two variance formulas need modification; the modified variances are available in Chao *et al.* (2015).

Both the Chao1 and Chao2 formulas were originally derived as lower bounds of species richness. Based on Alan Turing's statistical work, Chao *et al.* (2017) proved that when undetected species in the data have roughly the same mean relative abundance as singletons, the Chao1 lower bound becomes an unbiased estimator. Thus, a simple sufficient condition for the Chao1 lower bound being nearly unbiased is that *rare* species have approximately homogenous detection probability; in this case, the abundant species could be highly heterogeneous without affecting the estimator. Similar conditions can be applied to the Chao2 lower bound.

The Chao1 estimator may be very useful for data sets in which it is too time consuming to count the frequencies of all abundance classes, but it is relatively easy to count just the number of singleton and doubleton species. Chao1 and Chao2 are intuitive and very easy to calculate, and often perform just as well as more complex asymptotic estimators.

A more general approach is to use information on the frequency of other rare species, not just singletons and doubletons. A cut-off value κ denotes frequencies of rare species (frequency $\le \kappa$) and abundant species (frequency $> \kappa$). The cut-off $\kappa = 10$ works well with many empirical data sets. Let the total number of observed species in the abundant species group be $S_{\text{abun}} = \sum_{i>\kappa} f_i$ and the number of observed species in the rare species group be $S_{\text{rare}} = \sum_{i=1}^{\kappa} f_i$. Define $n_{\text{rare}} = \sum_{i=1}^{\kappa} i f_i$ and the coverage estimate $\hat{C}_{\text{rare}} = 1 - f_1/n_{\text{rare}}$. Coverage here is the estimated proportion of individuals in the rare species group that is represented by the species recorded in the sample; see the later subsection Coverage-based Standardization for details. The *Abundance-based Coverage Estimator* (*ACE*) is

$$\hat{S}_{\mathrm{ACE}} = S_{\mathrm{abun}} + \frac{S_{\mathrm{rare}}}{\hat{C}_{\mathrm{rare}}} + \frac{f_1}{\hat{C}_{\mathrm{rare}}}\hat{\gamma}^2_{\mathrm{rare}},$$

where $\hat{\gamma}^2_{\mathrm{rare}}$ is the square of the estimated coefficient of variation of the species relative abundances:

$$\hat{\gamma}^2_{\mathrm{rare}} = \max\left\{\frac{S_{\mathrm{rare}}}{\hat{C}_{\mathrm{rare}}}\frac{\sum_{i=1}^{\kappa} i(i-1)f_i}{(\sum_{i=1}^{\kappa} if_i)(\sum_{i=1}^{\kappa} if_i - 1)} - 1, \quad 0\right\}.$$

An approximate variance for the ACE can be obtained using a standard asymptotic approach or a bootstrap method.

For incidence data, there is a corresponding *Incidence-based Coverage Estimator* (ICE). As with ACE, we first select a cut-off point κ that partitions the data into an infrequent species group (incidence frequency not larger than κ) and a frequent species group (incidence frequency larger than κ). The cut-off $\kappa = 10$ is recommended. Denote the number of species in the frequent group by $S_{\mathrm{freq}} = \sum_{i>\kappa} Q_i$ and the number of species in the infrequent group by $S_{\mathrm{infreq}} = \sum_{i=1}^{\kappa} Q_i$. The estimated sample coverage for the infrequent group is $\hat{C}_{\mathrm{infreq}} = 1 - Q_1/\sum_{i=1}^{\kappa} Q_i$. Let the number of sampling units that include at least one infrequent species be R_{infreq}. Then ICE is expressed as

$$\hat{S}_{\mathrm{ICE}} = S_{\mathrm{freq}} + \frac{S_{\mathrm{infreq}}}{\hat{C}_{\mathrm{infreq}}} + \frac{Q_1}{\hat{C}_{\mathrm{infreq}}}\hat{\gamma}^2_{\mathrm{infreq}},$$

where $\hat{\gamma}^2_{\mathrm{infreq}}$ is the squared estimate of the coefficient of variation of the species relative incidences:

$$\hat{\gamma}^2_{\mathrm{infreq}} = \max\left\{\frac{S_{\mathrm{infreq}}}{\hat{C}_{\mathrm{infreq}}}\frac{R_{\mathrm{infreq}}}{(R_{\mathrm{infreq}}-1)}\frac{\sum_{i=1}^{\kappa} i(i-1)\,Q_i}{(\sum_{i=1}^{\kappa} iQ_i)(\sum_{i=1}^{\kappa} iQ_i - 1)} - 1, 0\right\}.$$

In addition to Chao1, Chao2, ACE, and ICE, the *jackknife* method (Burnham and Overton, 1979) provides another class of nonparametric estimators of asymptotic species richness. Jackknife techniques were developed as a general method to reduce the bias of a biased estimator. Here the biased estimator is the number of species observed in the sample. The basic idea with the jth order jackknife method is to consider data subsets by successively deleting j individuals from the data. The first-order jackknife turns out to be

$$\hat{S}_{jk1} = S_{\mathrm{obs}} + \frac{n-1}{n}f_1 \approx S_{\mathrm{obs}} + f_1.$$

That is, only the number of singletons is used to estimate the number of undetected species. The second-order jackknife estimator, which uses singletons and doubletons, has the form:

$$\hat{S}_{jk2} = S_{\mathrm{obs}} + \frac{2n-3}{n}f_1 - \frac{(n-2)^2}{n(n-1)}f_2 \approx S_{\mathrm{obs}} + 2f_1 - f_2.$$

Higher-order jackknife estimators, which use the more common species frequencies to estimate the number of undetected species, are available (Burnham and Overton, 1979). For replicated incidence data, jackknife estimators are obtained by replacing the sample size n in the above two formulas with the number of sampling units, and replacing abundance frequency counts with incidence frequency counts.

A basic assumption for the jackknife technique is that the bias of the initial estimator should be expressible as a power series in $1/n$. However, the observed species richness does not fulfill the assumption (Cormack, 1989, p. 404). Thus, the jackknife method does not have a theoretical basis for bias reduction in the context of species richness estimation. In addition, based on extensive simulation studies, Chiu *et al.* (2014b) found that the jackknife estimators often exhibit counter-intuitive patterns: their bias and accuracy do not improve as sample size increases, whereas the other nonparametric estimators presented above always improve. All the species richness estimators discussed above can be computed from the R package, SpadeR, available from CRAN. An online version is also available (Chao *et al.*, 2015).

Extrapolating Species Richness

Based on a reference sample of n individuals, the extrapolation or prediction problem is to estimate the expected number of species $S_{\mathrm{ind}}(n + m^*)$ in an augmented sample of $n + m^*$ individuals from the assemblage ($m^* > 0$). Under a simple multinomial model, Shen *et al.* (2003) derived the following useful predictor with an asymptotic variance:

$$\tilde{S}_{\mathrm{ind}}(n+m^*) = S_{\mathrm{obs}} + \hat{f}_0\left[1 - \left(1 - \frac{f_1}{n\hat{f}_0}\right)^{m^*}\right]$$

$$\approx S_{\mathrm{obs}} + \hat{f}_0\left[1 - \exp\left(-\frac{m^*}{n}\frac{f_1}{\hat{f}_0}\right)\right],$$

where $\hat{f}_0$ is an estimator for f_0 (the number of undetected species) based on either the Chao1 estimator ($\hat{f}_0 = \hat{S}_{\mathrm{Chao1}} - S_{\mathrm{obs}}$) or the ACE estimator ($\hat{f}_0 = \hat{S}_{\mathrm{ACE}} - S_{\mathrm{obs}}$).

The corresponding extrapolation formula and its asymptotic variance for the Coleman area-based Poisson sampling model was developed by Chao and Shen (2004). An estimator for the expected number of species $S_{\mathrm{area}}(A + a^*)$ in an augmented area $A + a^*$

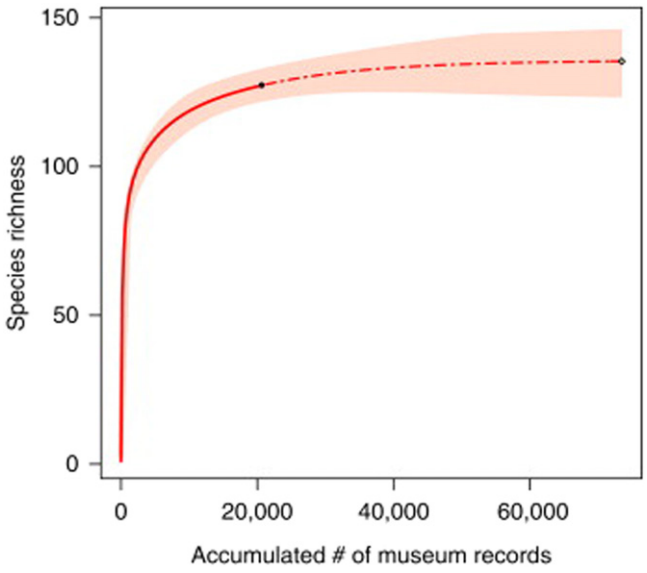

Fig. 5 A smoothed rarefaction and extrapolation curve. The *x*-axis is the number of individual, geo-referenced, dated ant specimens in New England, and the *y*-axis is the observed number of species. The total collection (the reference sample, filled circle) included 127 species and 20,225 individual records. The solid curve is the rarefaction curve interpolated from the reference sample. The dashed curve is the extrapolation, which extends to a minimum asymptotic estimator (Chao1) of ~135 species (open diamond). This number accords well with an independent estimate of an additional eight species that occur in suitable habitat in New York and Quebec. These eight species are likely to occur in New England, but so far they have not been collected there. However, the extrapolation to reach the Chao1 estimator extends to over 70,000 museum records, and the confidence interval (shaded polygon) is therefore fairly broad. The data set was compiled from museum records and private collections of ants sampled throughout the New England states of the USA (RI, CT, MA, VT, NH, and ME) between 1900 and 2011. Data modified from Ellison A. M., Gotelli, N.J., Farnsworth, E.J. and Alpert, G. D. (2012) A field guide to the ants of New England. New Haven, CT: Yale University Press.

($a^* > 0$) based on a reference sample of area A is

$$\tilde{S}_{\text{area}}(A + a^*) = S_{\text{obs}} + \hat{f}_0 \left[1 - \exp\left(-\frac{a^* f_1}{A \hat{f}_0}\right)\right],$$

where $\hat{f}_0$ is the same as in the individual-based model.

For incidence-based data with R sampling units comprising the reference sample, Colwell *et al.* (2012) developed a Bernoulli-product model and derived the following estimator for the expected number of species $S_{\text{sample}}(R + r^*)$ in an augmented set of $R + r^*$ sampling units ($r^* > 0$) from the assemblage:

$$\tilde{S}_{\text{sample}}(R + r^*) = S_{\text{obs}} + \hat{Q}_0 \left[1 - \left(1 - \frac{Q_1}{Q_1 + R\hat{Q}_0}\right)^{r^*}\right]$$

$$\approx S_{\text{obs}} + \hat{Q}_0 \left[1 - \exp\left(\frac{-r^* Q_1}{Q_1 + R\hat{Q}_0}\right)\right].$$

Here $\hat{Q}_0$, which is an estimator for Q_0, can be obtained from either the Chao2 estimator ($\hat{Q}_0 = \hat{S}_{\text{Chao2}} - S_{\text{obs}}$) or the ICE estimator ($\hat{Q}_0 = \hat{S}_{\text{ICE}} - S_{\text{obs}}$).

For each of these models, Colwell *et al.* (2012) linked the interpolation (rarefaction) curve and the corresponding extrapolation (prediction) curve to yield a single smooth curve meeting at the reference sample (**Fig. 5**). They also derived 95% (unconditional) confidence intervals for the interpolated and extrapolated richness estimates. Thus, rigorous statistical comparison can be performed not only for rarefaction but also for extrapolated richness values. This link helps to avoid the problem of discarding data and information from larger samples that is necessary for comparisons using the traditional rarefaction method. However, the extrapolations become highly uncertain if they are extended beyond approximately double the reference sample size. For both individual- and sample-based data, the additional sample size needed, beyond the reference sample, to attain the estimated asymptotic species richness, or to detect a specified proportion of asymptotic richness, is provided in Chao *et al.* (2009) and Colwell *et al.* (2012).

Species/Taxonomic Diversity

Species Diversity Metrics

Although species richness is the most popular and intuitive measure for characterizing diversity, the section Species Richness Estimation emphasizes that it is a very difficult parameter to estimate reliably from small samples, especially for hyperdiverse assemblages with many rare species. Species richness also does not account for the evenness of the species abundance distribution. Over the span of many decades, ecologists have proposed a plethora of diversity measures that incorporate both species richness and evenness, using both parametric and nonparametric approaches (Magurran, 2004).

Parametric approaches assume a particular species abundance distribution (such as the log-normal or gamma) or a species rank abundance distribution (such as the negative binomial or log series), and then estimate parameters from the distribution model that quantify the heterogeneity among species in their relative frequencies. For example, in the Fisher *et al.* (1943) pioneering sampling model, species rank abundances follow a log-series distribution. One of the parameters characterizing the log-series is the well-known Fisher's "alpha," which has been widely used to characterize ecological data as a diversity index. However, Fisher's alpha only uses the information of observed species richness and the total number of individuals; species abundances are ignored. Parametric models usually do not perform well unless the "true" species abundance distribution is known, which is never the case (Colwell and Coddington, 1994).

Nonparametric methods make no assumptions about the mathematical form of the underlying species abundance distribution. Scientists in other disciplines developed a wide range of "complexity" measures (which consider both evenness and richness) much earlier than ecologists. For example, a popular complexity measure is the *Shannon entropy*, developed by Claude Shannon in 1948 in information science:

$$H_{\mathrm{Sh}} = -\sum_{i=1}^{S} p_i \log p_i,$$

where S is the number of species in the assemblage and the i-th species relative abundance is p_i. Shannon entropy quantifies the uncertainty in the species identity of a randomly chosen individual in the assemblage. Ramón Margalef introduced Shannon entropy to ecology as a diversity index around 1957.

Another popular measure is the *Gini-Simpson index*, which was originally developed by the Italian statistician and sociologist Corrado Gini around 1912:

$$H_{\mathrm{GS}} = 1 - \sum_{i=1}^{S} p_i^2.$$

This index represents the probability that two randomly chosen individuals (selected with replacement) belong to two different species. In 1949, Simpson introduced the complement measure ("Simpson index" or concentration index) to ecology (Simpson, 1949). With an adjustment for z_+ (the total number of individuals in the assemblage), the Gini–Simpson index is closely related to the ecological index *PIE* (Hurlbert, 1971), the probability of an interspecific encounter:

$$\mathrm{PIE} = [z_+/(z_+ - 1)]H_{\mathrm{GS}},$$

which measures the probability that two randomly chosen individuals (selected *without* replacement) belong to two different species. Both PIE and the Gini–Simpson index have a straightforward interpretation as a probability. When PIE is applied to species abundance data, it is equivalent to the slope of the individual-based rarefaction curve measured at its base.

However, the units of the Gini–Simpson index and PIE are probabilities that are bounded between 0 and 1, and the units of Shannon entropy are logarithmic units of information. As a consequence, these popular diversity measures do not behave in the same intuitive way as species richness (Jost, 2007). The ecologist MacArthur (1965) was the first to show that Shannon entropy (when computed using natural logarithms) can be transformed to its exponential $\exp(H_{\mathrm{Sh}})$, and the Gini–Simpson index can be transformed to $1/(1 - H_{\mathrm{GS}})$, yielding two new indices that measure diversity in units of species richness. In particular, these transformed indices measure diversity in units of "effective number of species"—the equivalent number of equally abundant species that would be needed to give the same value of the diversity measure. When all species are equally abundant, the effective number of species is equal to the species richness of the assemblage.

These converted measures, like species richness itself, satisfy an important and intuitive property called the "replication principle" or the "doubling property" (Hill, 1973): if N equally diverse assemblages with no shared species are pooled in equal proportions, then the diversity of the pooled assemblages should be N times the diversity of each single assemblage. Simple examples show that Shannon's entropy and the Gini–Simpson measures do not obey the "replication principle." In contrast, the transformed values of these indices do obey the replication principle.

Hill Numbers

The ecologist Mark Hill incorporated the transformed Shannon and Gini–Simpson measures, along with species richness, into a family of diversity measures later called "Hill numbers," all of which measure diversity as the effective number of species. As defined earlier, let z_i denote the absolute abundance of species i, z_+ denote total abundance, and $p_i = z_i/z_+$ denote species relative

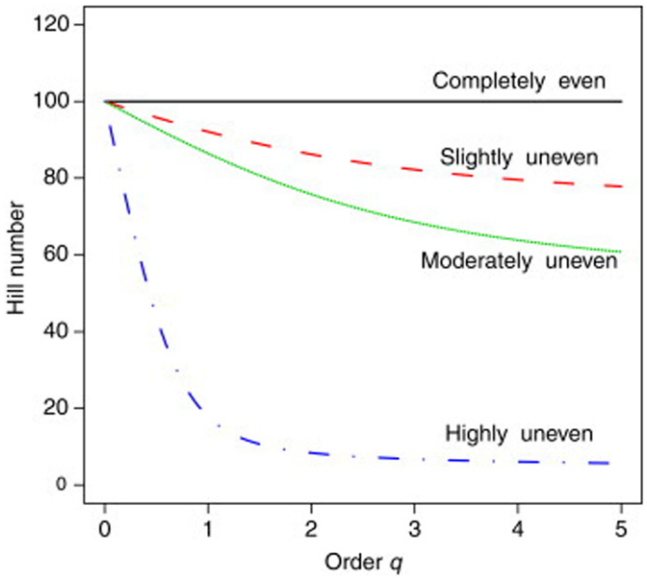

Fig. 6 Diversity profile for assemblages of differing evenness. The *x*-axis is the order q in the Hill number (Eqs. (2a) and (2b)), and is illustrated for values of q from 0 to 5. The *y*-axis is the calculated Hill number (the equivalent number of equally abundant species). Each of the four assemblages has exactly 100 species and 500 individuals, but they differ in their relative evenness: (1) completely even assemblage (black solid line): each species is represented by five individuals; (2) slightly uneven assemblage (red dashed line): 50 species each represented by seven individuals and 50 species each represented by three individuals (this structure is denoted as {50 × 7, 50 × 3}); (3) moderately uneven assemblage (green dotted line): {22 × 10, 28 × 5, 40 × 3, 10 × 2}; (4) highly uneven assemblage (blue dash–dot line): {1 × 120, 1 × 80, 1 × 70, 1 × 50, 3 × 20, 3 × 10, 90 × 1}. For $q = 0$, the Hill number is species richness, which is equal to 100 for all assemblages. Because Hill numbers represent the equivalent number of equally abundant species, the curve for the perfectly even assemblage (black solid line) does not change as q is increased. Larger values of q place progressively more weight on common species, so the equivalent number of equally abundant species is much lower for the more uneven assemblages than for more even assemblages.

abundance. Different Hill numbers ${}^qD$ are defined by their "order" q (Hill, 1973)

$$ {}^qD = \left(\sum_{i=1}^{S} \left(\frac{z_i}{z_+} \right)^q \right)^{1/(1-q)} = \left(\sum_{i=1}^{S} p_i^q \right)^{1/(1-q)}, \quad q \geq 0, \quad q \neq 1. \tag{2a} $$

This equation is undefined for $q = 1$, but in the limit as q tends to 1:

$$ {}^1D = \lim_{q \to 1} {}^qD = \exp\left(- \sum_{i=1}^{S} p_i \log p_i \right) = \exp\left(H_{\mathrm{sh}} \right). \tag{2b} $$

The parameter q controls the sensitivity of the measure to species relative abundance. When $q = 0$, the species relative abundances do not count at all, and ${}^0D$ equals species richness. When $q = 1$, the Hill number ${}^1D$ (Shannon diversity) is the exponential form of Shannon entropy, which weights species in proportion to their frequency and can be roughly interpreted as the number of "abundant species" in the assemblage. When $q = 2$,

$$ {}^2D = 1 / \sum_{i=1}^{S} p_i^2 = 1/(1 - H_{\mathrm{GS}}), $$

which heavily weights the most common species in the assemblage; the contribution from rare species is severely discounted. The measure ${}^2D$ (Simpson diversity) can be interpreted as the number of "very abundant species" in the assemblage. Because all Hill numbers of higher order place increasingly greater weight on the most abundant species, they are much less sensitive to sample size (number of individuals or sampling units surveyed) than the most commonly used Hill numbers ($q = 0, 1, 2$). Hill numbers with negative exponents can also be calculated, but they weight rare species heavily and thus have poor sampling properties.

The measure of diversity using Hill numbers can potentially depend on the order q that is chosen. However, because the order q can be any number (it need not be an integer), and all Hill numbers share a common unit of species richness, they can be portrayed on a single graph as a function of $q \geq 0$. This "diversity profile" of effective species richness versus q conveys all of the information about species abundance distribution of an assemblage (**Fig. 6**) and completely characterizes species/taxonomic diversity. The diversity profile curve is a decreasing function of q (Hill, 1973). The more uneven the distribution of relative abundances, the more steeply the curve declines. For a perfectly even assemblage, the profile curve is a constant at the level of species richness.

For incidence or occurrence data among multiple sampling units, Chao *et al.* (2014) defined the corresponding species diversity of order q as the Hill number based on species *relative detection* probabilities for any occurrence record. That is, the probability p_i in Eqs. (2a) and (2b) is replaced by the relative detection probability of species i in any sampling unit (i.e., the probability that any detected occurrence/incidence is classified to species i). Hill numbers for incidence data quantify the effective number of equally frequent species. For $q = 0$, the measure reduces to species richness, and the measures of $q = 1$ and $q = 2$ can be respectively interpreted as the effective number of frequent and highly frequent species in the assemblage. In the following, we mainly focus on abundance data, but all properties and inferences for abundance data can be readily extended to incidence data (Chao and Colwell, 2017).

Controlling for Sampling Effect

Empirical diversity analyses are always based on a limited amount of sample data, and species diversity is estimated by substituting species sample proportions into the formula of Hill numbers. When sampling is not complete, empirical Hill numbers of any order q are sensitive not only to shape of the species abundance distribution (SAD) but also to sample size and sample completeness. Although the shape of the SAD is an important biological attribute of an assemblage, sample size and sample completeness do not reflect biological processes, and their effects have to be controlled for when comparing the diversity of different samples.

For incomplete data, the undersampling effect and the SAD effect are confounded and cannot be easily disentangled. To rigorously compare diversity across assemblages, two procedures are recommended: (1) An asymptotic analysis, which removes the sampling effect via asymptotic diversity estimation (Chao and Jost, 2015). (2) A non-asymptotic analysis, which controls for sampling effect by standardizing sample size/completeness via the interpolation and extrapolation (iNEXT) method (Colwell *et al.*, 2012; Chao *et al.*, 2014), as detailed in later sections. An R package iNEXT, available from CRAN (Hsieh *et al.*, 2016), was developed to facilitate all computations and graphics. For users without an R background, an online version is also available; see "Relevant Websites" section.

Asymptotic Analysis: Asymptotic Diversity Estimates of Hill Numbers

An asymptotic analysis aims to compare asymptotic estimates of true diversities of entire assemblages. Here an "asymptotic" value refers to the diversity estimate that would be reached when the sample size is hypothetically extended to be very large; beyond this point, additional samples do not add any new species and do not change the relative proportions of the species that are already present in the sample. Based on the successive slopes of the species accumulation curve, Chao and Jost (2015, their Eq. 7d) derived an analytic, continuous diversity profile for *any* order $q \geq 0$ for both abundance and incidence data. A bootstrap method is applied to obtain the associated confidence intervals. For the special case of $q = 0$, the estimated species richness reduces to the richness estimators Chao1 (for abundance data) or Chao2 (for incidence data) richness estimator. For both abundance and incidence data, the asymptotic diversity estimator for $q = 1$ can substantially reduce undersampling bias, and the asymptotic estimate for $q = 2$ yields a corresponding nearly unbiased estimator.

Generally, Hill numbers of $q \geq 1$ for an entire assemblage can be accurately estimated with negligible bias, if samples are not too small. However, for Hill numbers of $q < 1$, especially for species richness ($q = 0$), asymptotic diversity estimates typically are subject to negative bias (they are underestimates) and represent only lower bounds. Whether data are sufficient to accurately infer the true diversities based on the asymptotic estimates can be determined by visually examining the size-based rarefaction and extrapolation sampling curve; see below.

Non-Asymptotic Analysis: The iNEXT Standardization

When samples are too small to provide accurate asymptotic diversity estimates, a non-asymptotic approach is recommended. This approach aims to compare diversity estimates for equally large or equally complete samples via rarefaction and extrapolation, under a framework of Hill numbers.

Size-based standardization

This standardization represents the traditional comparative approach in ecology (Gotelli and Colwell, 2011). Here, sample "size" typically means the number of individuals sampled for abundance data, or the number of sampling units for incidence data. In this standardization, all samples are compared at a common hypothetical sample size, which can be less than the reference sample size (rarefaction) or greater than the reference sample size (extrapolation). For the special case of $q = 0$, all diversity formulas for rarefied samples and extrapolated samples reduce to those for species richness given in the section Species Richness Estimation. The diversity estimate is thus depicted as a function of sample size up to a maximum size. For species richness, the estimate can generally be extrapolated to at most double the observed sample size. For Shannon diversity and Simpson diversity, if data are not sparse, the extrapolation can be reliably extended to the asymptote.

The pattern of size-based sampling curves can be used to visually determine whether the data contain sufficient information to infer the diversities of entire assemblages based on the asymptotic estimates. If the curve becomes stable and levels off when size is extrapolated to double the reference sample size, then the bias of the asymptotic estimator

is limited or negligible and thus the true diversity can be inferred from the sampling data. In contrast, if the curve is still increasing (this typically occurs for species richness, $q = 0$), it signifies that the asymptotic species-richness estimator represents only a lower bound. In this case, the following coverage-based approach is suggested.

Coverage-based standardization

Chao and Jost (2012) proposed comparing assemblages by plotting their diversities as a function of sample completeness. The measure used to quantify sample completeness is what we refer to as sample coverage (or simply "coverage"), i.e., the fraction of individuals in the entire assemblage that belong to detected species. The concept of sample coverage was originally developed by Alan Turing and I. J. Good in their cryptographic analysis during World War II (Good, 1953) and has subsequently been applied in various disciplines. Sample coverage can be very efficiently and accurately estimated directly from sampling data for the reference sample, rarefied samples, and extrapolated samples.

In coverage-based standardization, all samples are standardized by their sample coverage, which can be either smaller than the coverage of the observed sample size (rarefaction) or larger (extrapolation). The seamless rarefaction and extrapolation curve plots the diversity estimate as a function of sample coverage up to a maximum coverage. For species richness, the maximum coverage for comparing assemblages is specified to be the coverage corresponding to double the reference sample size, for each sample. For Shannon diversity and Simpson diversity, if data are not sparse, the extrapolation can often be extended to the coverage of unity to attain the estimated asymptote.

The size- and coverage-based sampling curves can be connected in the following sense: the slope of the size-based curve for a hypothetical sample of any size is equal to the coverage deficit (one minus the coverage) of the sample. Thus, standardizing coverage is equivalent to standardizing the slope of sized-based sampling curves. Contrary to the traditional standardization, standardizing sample coverage (or slope) generally requires that sample size vary among assemblages. In coverage-based standardization, smaller samples are required for less diverse assemblages, whereas larger samples are required for more diverse assemblages. Coverage-based standardization is based on the assemblage's characteristics (i.e., slopes of the estimated species accumulation curve), rather than sample size determined by samplers or happenstance. When the focus is on assessing species diversity across more than one assemblage, the iNEXT standardization is first applied to each assemblage separately, followed by comparisons of standardized diversity across multiple assemblages.

Taxonomic (Dis)similarity

Classic Richness-Based (Dis)similarity Indices

The earliest published richness-based measure of species compositional similarity is the classic Jaccard index from 1900. The classic Sørensen similarity index was proposed in 1948. These two classic indices are the most widely used ones, and both were originally developed to compare the similarity of two assemblages. Let S_1 be the number of species in Assemblage 1, S_2 be the number of species in Assemblage 2, and S_{12} be the number of shared species. The Jaccard similarity index is $S_{12}/(S_1 + S_2 - S_{12})$, which compares the number of shared species to the total number of species in the combined assemblages. The Sørensen similarity index is $2S_{12}/(S_1 + S_2)$, which compares the number of shared species to the mean number of species in a single assemblage. The Jaccard index is thus a comparison relative to total (gamma) diversity, whereas the Sørensen index is a comparison relative to local (alpha) diversity. A rigorous definition of alpha, beta, and gamma diversity will be formulated in a later section.

The two classic similarity measures and the corresponding dissimilarity (one minus similarity) measures can be linked to Whittaker's (1960, 1972) richness-based beta diversity. In these pioneering papers, Whittaker first defined beta diversity as a measure of the extent of differentiation in species composition among a set of assemblages. He also formulated multiplicative beta diversity as a ratio of gamma (total number of species) and alpha (mean number of species per assemblage). Here, alpha diversity should not be confused with Fisher's alpha, although a close reading of Whittaker (1960) suggests that the latter inspired Whittaker's nomenclature; Whittaker (1960) also indicated that Fisher's alpha can be one measure of alpha diversity. For the special two-assemblage case, Whittaker's richness-based beta diversity is defined as $\beta_W = (S_1 + S_2 - S_{12})/[(S_1 + S_2)/2]$. Then the classic Jaccard and Sørensen dissimilarity indices can be expressed, respectively, as $2(1 - 1/\beta_W)$ and $\beta_W - 1$. Both the classic dissimilarity indices are thus monotonically increasing functions of Whittaker's richness-based beta diversity; the corresponding similarity indices are thus monotonically decreasing functions of Whittaker's richness-based beta diversity (Jost *et al.*, 2011).

When more than two assemblages are compared, a typical approach is to use the average of all pairwise similarities as a measure of global similarity. However, Chao *et al.* (2008) provided numerical examples in which all pairwise similarities are identical in two sets of assemblages, but the global similarities for the two sets are different. These examples illustrate that pairwise similarities cannot fully characterize multiple-assemblage similarity when some species are shared across two, three, or more assemblages.

In addition, the above two classic indices do not consider species abundance; abundant and rare species are treated equally, which oversimplifies the relationships between assemblages. If species abundances can be measured, they should be used for a more accurate representation (and better statistical estimation) of the similarity of assemblages. Since the 1950s, a wide range of (dis)similarity measures have been developed aiming to extend the two classic measures to incorporate species abundance and also to accommodate more than two assemblages. This article reviews those (dis)similarity measures under the framework of diversity decomposition (gamma, alpha, and beta) based on Hill numbers. In this framework, beta diversity can be transformed to

Table 1 A summary of gamma, alpha, and beta formulas for taxonomic diversity and phylogenetic diversity based on an S by N species abundance matrix $[z_{ik}]$, where N = the number of assemblages; S = the number of species in the pooled assemblage, $i = 1, 2, \ldots, S$, $k = 1, 2, \ldots, N$. For the phylogenetic diversity, z^*_{ik} and z^*_{i+} for $i = 1, 2, \ldots, B$ represent the total abundances descending from branch i, $i = 1, 2, \ldots, B$ in the k-th assemblage and the pooled assemblage, respectively. $z_{++} = \sum_{k=1}^{N}\sum_{i=1}^{S} z_{ik}$. Here z_{++} represents the total abundance in the matrix $[z_{ik}]$.

Diversity	*Taxonomic diversity (effective number of species)*	*Phylogenetic diversity (effective total branch length)*
Gamma diversity of order q	$^qD_\gamma = \left\{\sum_{i=1}^{S}\left(\frac{z_{i+}}{z_{++}}\right)^q\right\}^{1/(1-q)}$	$^qPD_\gamma = \left\{\sum_{i=1}^{B} L_i\left(\frac{z^*_{i+}}{Tz_{++}}\right)^q\right\}^{1/(1-q)}$ $mean\,^qPD_\gamma \equiv \overline{D}_\gamma = {}^qPD_\gamma/T$ (effective number of lineages)
Alpha diversity of order q	$^qD_\alpha = \frac{1}{N}\left\{\sum_{i=1}^{S}\sum_{k=1}^{N}\left(\frac{z_{ik}}{z_{++}}\right)^q\right\}^{1/(1-q)}$	$^qPD_\alpha = \frac{1}{N}\left\{\sum_{k=1}^{N}\sum_{i=1}^{B} L_i\left(\frac{z^*_{ik}}{Tz_{++}}\right)^q\right\}^{1/(1-q)}$ $mean\,^qPD_\alpha \equiv {}^q\overline{D}_\alpha = {}^qPD_\alpha/T$
Beta diversity of order q	$^qD_\beta = \frac{^qD_\gamma}{^qD_\alpha} = N\left(\frac{\sum_{i=1}^{S} z^q_{i+}}{\sum_{i=1}^{S}\sum_{k=1}^{N} z^q_{ik}}\right)^{1/(1-q)}$	$^qPD_\beta = \frac{^qPD_\gamma}{^qPD_\alpha} = {}^q\overline{D}_\beta = \frac{^q\overline{D}_\gamma}{^q\overline{D}_\alpha}$ $= N\left(\frac{\sum_{i=1}^{B} L_i (z^*_{i+})^q}{\sum_{i=1}^{B} L_i \sum_{k=1}^{N} (z^*_{ik})^q}\right)^{1/(1-q)}$

(dis)similarity measures. The resulting (dis)similarity metrics encompass not only the richness-based classic Jaccard and Sørensen indices, but also abundance-sensitive Horn and Morisita-Horn measures. Moreover, the framework can be readily extended to a corresponding phylogenetic version.

Taxonomic Diversity Decomposition (Gamma, Alpha and Beta Diversity)

Assume that, in a species pool of N assemblages, there are S species, indexed by 1, 2, ..., S. Let $[z_{ik} \geq 0]$ be an $S \times N$ raw abundance matrix, where z_{ik} represents species *raw abundance* (i.e., the true number of individuals or other abundance proxies) of the i-th species in the k-th assemblage, $i = 1, 2, \ldots, S$, $k = 1, 2, \ldots, N$. The total abundance in the matrix is denoted by $z_{++} = \sum_{k=1}^{N}\sum_{i=1}^{S} z_{ik}$ where the index i is summed over the tip nodes only. Depending on the comparison target, we could also let z_{ik} represent within-assemblage relative abundance, biomass, or spatial coverage of corals, etc.

Gamma diversity is defined as the diversity of the *pooled assemblage* in which the abundance of any species is obtained by directly pooling its abundances over the N assemblages. That is, given the raw abundance matrix $[z_{ik}]$, the pooled assemblage comprises S species and the species abundance vector is $(z_{1+}, z_{2+}, \ldots, z_{s+})$, where $z_{i+} = \sum_{k=1}^{N} z_{ik}$ denotes the total abundance of species i in the pooled assemblage. Thus, gamma diversity requires only information about the species identification/classification for each individual; assemblage affiliation is not involved. Gamma diversity at the assemblage level is computed as Hill numbers for the pooled assemblage; see **Table 1** (middle column) for the gamma formula. Gamma diversity is interpreted as the effective number of species in the pooled assemblage.

Although the formulation of gamma diversity is unequivocal, there have been at least three different formulations of alpha diversity proposed for N assemblages in the ecological literature. In conventional formulations, alpha diversity is expressed as a weighted mean of the diversities of individual assemblages. Here, the formulation of Chiu *et al.* (2014a) is adopted; see **Table 1** (middle column) for the alpha formula, in which each individual is associated with two classifications—species identity and assemblage affiliation. The $S \times N$ cells of the raw abundance matrix are treated as if each cell were a "species." Each cell represents a species-assemblage combination. Alpha diversity is interpreted as the effective number of species-assemblage combinations per assemblage. If all assemblages have the same size, alpha diversity reduces to the average effective number of species per assemblage.

Based on the gamma and alpha formulas, beta diversity is defined as the ratio of gamma to alpha diversities. That is, we apply Whittaker's original multiplicative definition of beta diversity but use Hill numbers for any diversity order $q \geq 0$. Beta diversity is then expressed in units of "assemblage equivalents" or effective number of assemblages. Beta diversity

Table 2 A summary of four classes of taxonomic and phylogenetic compositional dissimilarity measures with interpretations (all are interpreted in an effective sense) based on monotonic transformations of beta diversity. Because beta diversity based on ${}^qPD$ and ${}^q\overline{D}$ are identical (i.e., ${}^qPD_\beta = {}^q\overline{D}_\beta$; see **Table 1**), all the ${}^q\overline{D}$ terms in the right column can be replaced by ${}^qPD$

Dissimilarity measure	*Monotonic transformations of taxonomic beta diversity*	*Monotonic transformations of phylogenetic beta diversity*
Sørensen-type non-overlap	$1-C_{qN}=\frac{({}^qD_\beta)^{1-q}-1}{N^{1-q}-1}$ $=\frac{\sum_{i=1}^{S}\sum_{k=1}^{N}(z_{ik}^q-\overline{z}_{i+}^q)}{(N^{1-q}-N^{2-2q})\ z_{++}^q({}^qD_\alpha)^{1-q}}$ Effective average proportion of non-shared *species* in an assemblage	$1-\overline{C}_{qN}=\frac{({}^q\overline{D}_\beta)^{1-q}-1}{N^{1-q}-1}$ $=\frac{\sum_{i=1}^{B}L_i\sum_{k=1}^{N}\left[(z_{ik}^*)^q-(\overline{z}_{i+}^*)^q\right]}{(N^{1-q}-N^{2-2q})\ z_{++}^q({}^qPD_\alpha)^{1-q}}$ Effective average proportion of non-shared *lineages* in an assemblage
Jaccard-type non-overlap	$1-U_{qN}=\frac{({}^qD_\beta)^{q-1}-1}{N^{q-1}-1}$ $=\frac{\sum_{i=1}^{S}\sum_{k=1}^{N}(z_{ik}^q-\overline{z}_{i+}^q)}{(1-N^{1-q})\ z_{++}^q({}^qD_\gamma)^{1-q}}$ Effective proportion of non-shared *species* in the pooled assemblage	$1-\overline{U}_{qN}=\frac{({}^q\overline{D}_\beta)^{q-1}-1}{N^{q-1}-1}$ $=\frac{\sum_{i=1}^{B}L_i\sum_{k=1}^{N}\left[(z_{ik}^*)^q-(\overline{z}_{i+}^*)^q\right]}{(1-N^{1-q})\ z_{++}^q({}^qPD_\gamma)^{1-q}}$ Effective proportion of non-shared *lineages* in the pooled assemblage
Sørensen-type turnover	$1-V_{qN}=\frac{{}^qD_\beta-1}{N-1}=\frac{{}^qD_\gamma-{}^qD_\alpha}{(N-1)({}^qD_\alpha)}$ $(V_{0N}=C_{0N};V_{2N}=U_{2N})$ *Species* turnover relative to alpha	$1-\overline{V}_{qN}=\frac{{}^q\overline{D}_\beta-1}{N-1}=\frac{{}^q\overline{D}_\gamma-{}^q\overline{D}_\alpha}{(N-1)({}^q\overline{D}_\alpha)}$ $(\overline{V}_{0N}=\overline{C}_{0N};\overline{V}_{2N}=\overline{U}_{2N})$ *Lineage* turnover relative to alpha
Jaccard-type turnover	$1-S_{qN}=\frac{1/{}^qD_\beta-1}{1/N-1}=\frac{{}^qD_\gamma-{}^qD_\alpha}{(1-1/N)({}^qD_\gamma)}$ $(S_{0N}=U_{0N};S_{2N}=C_{2N})$ *Species* turnover relative to gamma	$1-\overline{S}_{qN}=\frac{1/{}^q\overline{D}_\beta-1}{1/N-1}=\frac{{}^q\overline{D}_\gamma-{}^q\overline{D}_\alpha}{(1-1/N)({}^q\overline{D}_\gamma)}$ $(\overline{S}_{0N}=\overline{U}_{0N};\overline{S}_{2N}=\overline{C}_{2N})$ *Lineage* turnover relative to gamma

always attains a *fixed minimum* value of unity if all N communities are identical in terms of class identity and abundance, and beta diversity attains a *fixed maximum* value of N (the number of assemblages) when no species are shared among assemblages (i.e., complete turnover).

There are three reasons for using the alpha formula of Chiu *et al.* (2014a): (1) When $q = 0$, the beta formula reduces to Whittaker's original richness-based definition. (2) The Chiu *et al.* (2014a) formula avoids the problem that other measures of alpha diversity sometimes yield a beta value less one (i.e., gamma < alpha) or greater than N. (3) Chao and Chiu (2016) showed that only through this alpha formulation can a bridge be built between the two major approaches to beta diversity, i.e., the variance framework (Legendre and De Cáceres, 2013) and the diversity decomposition approach (presented here).

Taxonomic (Dis)similarity Measures

Because beta diversity in the Hill number framework lies in the range of $[1, N]$, it can be compared across studies only for a fixed value of N. To illustrate with a simple example, a beta value of 2 would signify maximal differentiation (no shared species) between two assemblages ($N = 2$), whereas the same beta value for a 20-assemblage case ($N = 20$) signifies very low differentiation. When N varies across studies, proper normalizations are needed to transform beta diversity to dissimilarity measures in a fixed range of $[0,1]$ by means of monotonic transformations; see **Table 2**. The resulting (dis)similarity measures can be compared even if N varies across studies. Here we briefly review, from two perspectives, the four classes of (dis)similarity measures which encompass the most widely used dissimilarity measures as special cases. Beta diversity and the associated dissimilarity measures of $q = 0$ reflect how species identities shift, whereas the corresponding measures of $q > 0$ reflect how species abundance distributions change among assemblages.

A perspective on measuring species composition turnover (or its complement)

The *turnover* of species composition is defined as a function of diversity excess (${}^qD_\gamma - {}^qD_\alpha$) between the pooled assemblage (on a regional scale; gamma) and an average assemblage (on a local scale; alpha). For example, Whittaker (1972) used $({}^qD_\gamma - {}^qD_\alpha)/{}^qD_\alpha$ (for the special case of $q = 0$) as a species turnover relative to alpha, which attains the minimum value of 0

when the N assemblages are identical and attains the maximum value of $(N-1)$ when the N assemblages share no species. Transforming this relative species turnover to the range [0,1] leads to the normalized measure $1 - V_{qN}$ (simply a linear function of the beta metric) in **Table 2**. Thus, the dissimilarity measure $1 - V_{qN}$ quantifies the normalized species (or species-equivalents) turnover rate with respect to the average assemblage (i.e., alpha). Similarly, considering the relative excess with respect to gamma leads to the class of Jaccard-type turnover measures $1 - S_{qN}$, which quantifies the normalized species turnover rate with respect to the pooled assemblage (i.e., gamma). This measure is a linear function of the inverse-beta metric.

A perspective on measuring species compositional overlap (or non-overlap)

Chao and Chiu (2016) showed that two major approaches to beta diversity (i.e., variance framework and diversity decomposition approach) lead to the same two classes of Jaccard- and Sørensen-type (non)overlap measures. The Sørensen-type overlap measure C_{qN} (or its complement $1 - C_{qN}$) quantifies the effective average proportion of shared species (or non-shared for its complement) in an individual assemblage. The Jaccard-type overlap measure U_{qN} (or its complement $1 - U_{qN}$) quantifies the effective proportion of shared species (or non-shared for its complement) in the pooled assemblage. This class of dissimilarity measures yield non-linear functions of inverse-beta diversity, if $q > 0$.

The four classes of (dis)similarity measures in **Table 2** for $q = 0$, 1, and 2 include several widely used metrics as special cases. For the special case of $q = 0$, richness-based beta diversity reduces to Whittaker's richness-based formula ${}^0D = \mathrm{S}/\overline{\mathrm{S}}$, a ratio of S (total species richness) to the simple mean species richness. The two classic richness-based Jaccard and Sørensen dissimilarity indices are then extended to N assemblages:

$S_{0N} = U_{0N} = \frac{\overline{S}/S - 1/N}{1 - 1/N}$ (richness-based N-assemblage Jaccard similarity);

$C_{0N} = V_{0N} = \frac{N - S/\overline{S}}{N-1}$ (richness-based N-assemblage Sørensen similarity).

For $N = 2$, the above two similarity measures reduce to the classic richness-based Jaccard and Sørensen two-assemblage similarity indices.

When the element of the data matrix $[z_{ik}]$ represents the relative abundance of species i within the k-th assemblage, the Sørensen- and Jaccard-type non-overlap measures of $q = 1$ are identical, i.e., $1 - C_{1N} = 1 - U_{1N} = (H_\gamma - H_\alpha)/\log N$, where H_γ and H_α denote gamma and alpha entropy. This dissimilarity measure is the well-known Shannon differentiation or Horn's (1966) dissimilarity measures for comparing equally-weighted species relative abundance sets based on Shannon entropy. In addition, for the special case of $q = 2$, the measure C_{2N} reduces to the traditional N-assemblage Morisita-Horn (Morisita, 1959) similarity measure, and the measure U_{2N} reduces to the regional species-overlap measure (Chiu *et al.*, 2014a). These two measures of $q = 2$ are based on decomposing Gini-Simpson index.

Controlling for Sampling Effects for Beta and (Dis)similarity Measures

Like all diversity measures, beta diversity and its associated (dis)similarity measures are estimated from sampling data. The classic richness-based Jaccard and Sørensen similarity indices are notoriously sensitive to sample size and sample completeness, especially for assemblages with numerous rare species. Thus, asymptotic estimation and non-asymptotic rarefaction and extrapolation are also required.

The observed classic Jaccard and Sørensen dissimilarity measures based on limited sampling generally over-estimate the true values because rare shared species may be undetected or detected only in one site and thus falsely recorded as unique species (Chao *et al.*, 2005). In the special case where N assemblages are identical, positive bias exists for richness-based dissimilarity even for very large sample sizes. On the other hand, beta diversity and dissimilarity values could also *under-estimate* when N assemblages share only very abundant species, while the rare ones tend to be locally endemic. In this case, two samples might yield the same few common species, but fail to reveal rare species that would differentiate the assemblages in larger samples. The asymptotic and non-asymptotic standardization procedures for gamma, alpha, beta, and the associated (dis)similarity measures follow from those for Hill numbers. For gamma diversity, all standardization procedures are performed based on the observed data in the pooled assemblage; for alpha diversity, all procedures are performed based on the observed two-dimensional (species-by-assemblage) data. Then the corresponding standardized beta and dissimilarity can be obtained; see Chao *et al.* (2023) for details. An illustrative example is shown later.

Phylogenetic Diversity and its Decomposition

Quantifying Phylogenetic Diversity

One major advantage of using Hill numbers is that the framework can be directly generalized to incorporate the evolutionary history among species. Assume that all species of a focal assemblage are connected by a rooted ultrametric phylogenetic tree (i.e., all branch tips are the same distance from the basal node), with the S species as tip nodes (**Fig. 2**). The base (or root) of the tree represents the most recent ancestral taxon, the branching forks (nodes) represent speciation or divergence events, and time is measured in the vertical axis, increasing from the base of the tree to the branch tips. A Linnean taxonomic tree, in which species are

simply classified into a taxonomic hierarchy (Kingdom, Phylum, Class, Order, Family, Genus, and Species), can be regarded as a special case of an ultrametric tree. All other things being equal, an assemblage in which all the species are closely related and concentrated in one region of the tree should be less diverse than an assemblage in which the same number of species is widely distributed among distant branch tips of the tree.

Assume that all phylogenetic diversity measures are computed from the age of the root, T, although the reference time point at any descendant node could also be used. Within a tree with depth T, assume that there are B branches/nodes in the phylogenetic tree, $B \geq S$. Let L_i denote the length of the i-th branch. For $i = 1, 2, \ldots, B$, define z_i^* (and p_i^*) as the total absolute (and relative) abundance descended from the i-th node/branch. That is, the set of species abundances is expanded to the set of node/branch abundances $\{z_i^*, i = 1, 2, \ldots, B\}$ with species abundances $(z_1, z_2, \ldots, z_S)$ as the first S elements. Likewise, the set of species relative abundances is expanded to the set of node/branch relative abundances $\{p_i^*, i = 1, 2, \ldots, B\}$, where $p_i^* = z_i^*/z_+$, $z_+ = \sum_{i=1}^{S} z_i$ (the index i in z_+ is summed over the tips only).

Pielou (1975) was the first to notice that the concept of diversity could be broadened to consider differences among species. The earliest diversity measure for a Linnean taxonomic tree is the *cladistic diversity* (*CD*), which is defined as the total number of taxa or nodes in a taxonomic tree that encompasses all of the species in the assemblage (Vane-Wright *et al.*, 1991). Another pioneering work is Faith's *phylogenetic diversity* (*PD*) (Faith, 1992), which is defined as the sum of the branch lengths of a phylogeny connecting all species in the target assemblage. In both *CD* and *PD*, neither species abundances nor node abundances are considered.

C.R. Rao's *quadratic entropy* (Q) was the first diversity measure that accounted for species abundances and species relatedness (Rao, 1982). It is a generalization of the Gini–Simpson index:

$$Q = \sum_{i,j=1}^{S} d_{ij}\, p_i\, p_j, \tag{3}$$

where d_{ij} denotes the phylogenetic (or functional) distance between species i and j, and p_i and p_j denote the relative abundance of species i and j. The phylogenetic or cophenetic distance of two species is defined as the average of the branch lengths connected to their most recent common ancestral node. This index Q measures the average phylogenetic (or functional) distance between any two individuals randomly selected from the assemblage. For the special case of no phylogenetic structure, all species are equally related to one another ($d_{ii} = 0$ and $d_{ij} = 1$ $(i \neq j)$), and Q reduces to the Gini–Simpson index. The *phylogenetic entropy* (H_P) is defined as a generalization of Shannon's entropy to incorporate phylogenetic distances among species (Allen *et al.*, 2009):

$$H_P = \sum_{i=1}^{B} L_i\, p_i^* \log p_i^*, \tag{4}$$

where the summation is over all branches.

The replication principle can be generalized to phylogenetic diversity: When N completely distinct trees (no shared nodes among trees, although nodes within a tree can be shared) with equal diversities are combined, the diversity of the combined tree is N times the diversity of any individual tree. As with their parent measures, neither Q nor H_P satisfies this replication principle, but can be transformed into measures that do obey the replication principle, as shown below.

Faith (1992) pointed out that there is an analogy between species and phylogenetic diversity. From Faith's perspective, each unit-length branch segment can be regarded as a phylogenetic "feature" or "attribute" (like "species"). This analogy to counting-up species means that most species diversity measures can be converted to PD equivalents by counting unit-length segments rather than species. Under this analogy, Chao *et al.* (2010) generalized Hill numbers to incorporate phylogeny information by considering a hypothetical assemblage that is decomposed into B sub-assemblages: the i-th sub-assemblage, $i = 1, 2, \ldots, B$, consists of L_i phylogenetic attributes (like "species" in Hill numbers), with all attributes having the same relative abundance $z_i^*/(Tz_+) = p_i^*/T$ (Note the total abundance is $\sum_{i=1}^{B} L_i z_i^* = Tz_+$ and thus a divisor Tz_+ is required to define the relative abundance). The Hill number of order q of the hypothetical assemblage is defined as the phylogenetic diversity of order q ($^qPD(T)$, or simply ^qPD):

$$^qPD = \left(\sum_{i=1}^{B} L_i \left(\frac{z_i^*}{Tz_+}\right)^q\right)^{1/(1-q)} = \left(\sum_{i=1}^{B} L_i \left(\frac{p_i^*}{T}\right)^q\right)^{1/(1-q)}, \quad q \geq 0, \quad q \neq 1. \tag{5}$$

For $q = 1$, ^1PD is defined as the limit of ^qPD when q tends to 1. The phylogenetic diversity ^qPD is interpreted as the *effective total branch length*, effective number of lineage-years, or effective total amount of evolutionary history in the phylogenetic tree with depth T. In **Fig. 7**, a simple example illustrates the major concepts and computation of ^qPD.

The *mean phylogenetic diversity of order q*, *mean* ^qPD or $^q\overline{D}$, is obtained by dividing the phylogenetic diversity by tree depth:

$$^q\overline{D} \equiv mean\,^qPD = \frac{^qPD}{T} = \left(\sum_{i=1}^{B} \frac{L_i}{T}\left(\frac{z_i^*}{z_+}\right)^q\right)^{1/(1-q)} = \left(\sum_{i=1}^{B} \frac{L_i}{T}\left(p_i^*\right)^q\right)^{1/(1-q)}, \quad q \geq 0, \quad q \neq 1. \tag{6}$$

This measure quantifies the effective number of equally abundant and equally divergent lineages each with lineage length T. Thus, this measure has the same units as the Hill numbers. If $T = 1$, then ^qPD is identical to $^q\overline{D}$. When there are no internal nodes, and the length of all lineages is T (equivalently, phylogeny is ignored), $^q\overline{D}$ reduces to the Hill number of order q. Both ^qPD (in units of "branch length" for all q) and $^q\overline{D}$ (in units of "lineages" for all q) satisfy the phylogenetic version of the replication principle; each

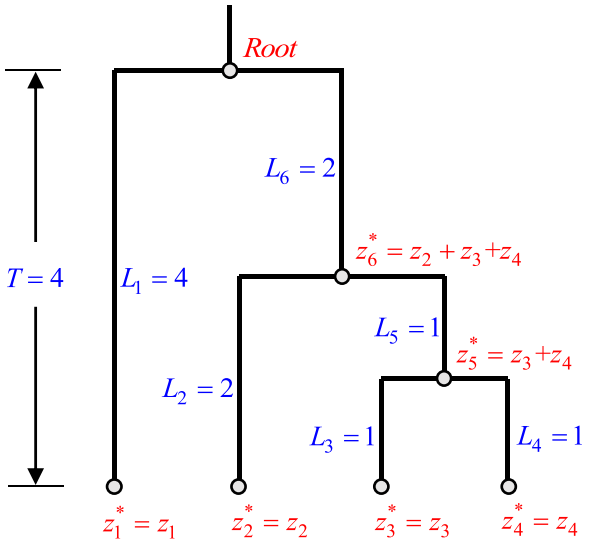

Fig. 7 A hypothetical ultrametric tree spanned by 4 extant species (tip nodes) indexed by 1, 2, 3, and 4, with raw abundances$(z_1, z_2, z_3, z_4) = (1, 2, 3, 4)$ and $z_+ = \sum_{i=1}^{4} z_i = 10$. The most recent ancestor of the four species is the root at the top, with time progressing towards the branch tips at the bottom. Here, the tree depth $= 4$ (i.e., the age of the root). The branch/node set in the tree includes $B = 6$ branches/nodes (4 tips and 2 internal nodes), indexed from 1 to 6, with branch lengths $(L_1, L_2, ..., L_6) = (4, 2, 1, 1, 1, 2)$ and the corresponding node/branch abundance set $(z_1^*, z_2^*, ..., z_6^*) = (z_1, z_2, z_3, z_4, z_3 + z_4, z_2 + z_3 + z_4) = (1, 2, 3, 4, 7, 9)$ and $\sum_{i=1}^{6} L_i z_i^* = Tz_+ = 40$. Conceptually, consider a hypothetical assemblage that is decomposed into $B = 6$ sub-assemblages: the first sub-assemblage consists of $L_1 = 4$ "species" (i.e., 4 unit-length branch segments) each with relative abundance $z_1^*/(Tz_+) = 1/40$; the second sub-assemblage consists of $L_2 = 2$ "species" each with relative abundance $z_2^*/(Tz_+) = 2/40$; the third sub-assemblage consists of $L_3 = 1$ "species" each with relative abundance $z_3^*/(Tz_+) = 3/40$,..., and the sixth sub-assemblage consists of $L_6 = 2$ "species" each with relative abundance $z_6^*/(Tz_+) = 9/40$. It then follows from Eq. (5) that the phylogenetic diversity of order q is simply the Hill number of order q for this hypothetical assemblage, i.e., $^qPD = \left(\sum_{i=1}^{B} L_i \times (z_i^*/Tz_+)^q\right)^{1/(1-q)} = [4 \times (1/40)^q + 2 \times (2/40)^q + 1 \times (3/40)^q + ... + 2 \times (9/40)^q]^{1/(1-q)}$.

measure for all orders $q \geq 0$ can be plotted on a single curve as a continuous function of the order $q \geq 0$. The phylogenetic q-profile fully conveys all information in the species abundance distribution and phylogenetic tree with depth T.

For $q = 0$, ^0PD reduces to Faith's PD, and $^0\overline{D}$ reduces to lineage richness. ^1PD and $^1\overline{D}$ of order $q = 1$ can be connected to phylogenetic entropy H_P (in Eq. (4)) by the following transformations: $^1PD = T\exp(H_P/T)$ and $^1\overline{D} = \exp(H_P/T)$. ^2PD and $^2\overline{D}$ of order $q = 2$ can be connected to quadratic entropy Q (in Eq. (3)) by the following transformations: $^2PD = T/(1 - Q/T)$ and $^2\overline{D} = 1/(1 - Q/T)$. The above simple transformations convert H_P and Q to measures that satisfy the phylogenetic version of the replication principle.

Because ^qPD (in Eq. (5)) represents genuine Hill numbers for a hypothetical assemblage stated above, the corresponding asymptotic estimation and non-asymptotic standardization for ^qPD and $^q\overline{D}$ can be formulated in a parallel way (Hsieh and Chao, 2017). The iNEXT standardization has also been expanded to phylogenetic and functional versions; see Chao *et al.* (2021) for details. A real-data example is given in a later section.

Phylogenetic Generalization of the two Classic (Dis)similarity Measures

Remarkable advances in high-throughput DNA sequencing and progress in molecular technology/tools have opened a novel way to study highly diverse microbial communities in various environments such as soils, oceans, animals and the tissues of the human body. Assessing species compositional and phylogenetic (dis)similarity among microbial communities has been one of the central issues in molecular biology and related sciences. Based on Faith's total branch lengths in a phylogenetic tree, microbiologists generalized the classic Jaccard and Sørensen measures to their phylogenetic versions. For example, Lozupone and Knight (2005) proposed the phylogenetic Jaccard dissimilarity measure, known as the *UniFrac* index, which quantifies the extent of evolutionary history that is unique to either of the two assemblages. The corresponding similarity index takes the form of $L_{12}/(L_1 + L_2 - L_{12})$, where L_1 and L_2 denote the total branch lengths in Assemblages 1 and 2, respectively, and L_{12} denotes the total length of the shared branches. Bryant *et al.* (2008) and Ferrier *et al.* (2007) proposed a classic phylogenetic Sørensen similarity index, known as *PhyloSør* index, which is expressed as $2L_{12}/(L_1 + L_2)$, In these two phylogenetic generalizations, "species" in the traditional taxonomic indices are replaced by the total branch lengths in each assemblage, and "shared species" in the traditional indices are replaced by the total shared branch length.

However, neither the UniFrac nor the PhyloSør index takes into account species or node abundance. In microbiology, a number of "quantitative" (dis)similarity measures that incorporate species and node abundance have been proposed but without theoretical

justification. This article reviews those phylogenetic (dis)similarity measures via phylogenetic diversity partitioning theory under the framework of Hill numbers (see the right columns in **Tables 1** and **2**). The resulting (dis)similarity metrics encompass some of the existing (dis)similarity measures based on decomposing abundance-sensitive phylogenetic entropy (Eq. (4)) and Rao's quadratic entropy (in Eq. (3)). All procedures are parallel to those for taxonomic diversity, and thus most details are omitted; see Alberdi and Gilbert (2019) for a review specifically on DNA-based diversity analysis.

Phylogenetic Diversity Decomposition for *N* Assemblages

When there are N assemblages comprising a species pool of S species, all phylogenetic diversity measures should be computed from a fixed reference point in time. Assume a phylogenetic tree spanned by S species has a root of age T. Assume that there are B branches/nodes in the tree and let L_i denote the length of the i-th branch, $i = 1, 2, \ldots, B$. As in the one-assemblage case, for $i = 1, 2, \ldots, B$, the definition of z_{ik} is extended to include all internal nodes/branches by defining z^*_{ik} and z^*_{i+} for $i = 1, 2, \ldots, B$ as the total abundances descending from branch i in the k-th assemblage and the pooled assemblage, respectively.

Phylogenetic diversity and mean phylogenetic diversity can be decomposed in a parallel way as taxonomic diversity. All relevant gamma, alpha, and beta diversity formulas are shown in **Table 1**. The phylogenetic gamma diversity is interpreted as the effective total branch length (or total amount of evolutionary history) in the pooled assemblage. Phylogenetic alpha diversity can be interpreted as the effective total branch length per assemblage. The mean phylogenetic gamma and alpha diversity can be analogously interpreted in terms of the effective number of equally divergent and equally abundant lineages. Based on a multiplicative partitioning, both phylogenetic diversity and mean phylogenetic diversity lead to the same phylogenetic beta diversity.

The phylogenetic beta diversity is always between 1 and N, for *all* $q \geq 0$, regardless of species abundance and tree structure. When the N assemblages are identical in terms of species identity and species raw abundance, phylogenetic beta diversity attains a minimum value of 1. When the N assemblages are completely phylogenetically distinct (no shared branches across assemblages, though branches within any assemblage may be shared), phylogenetic beta diversity attains a maximum value of N. The phylogenetic beta component quantifies phylogenetic differentiation among multiple sets of node abundances and is interpreted as the effective number of completely phylogenetically distinct assemblages based on node abundances. When all branches in the pooled assemblage are unshared and T is standardized to unity, phylogenetic alpha, beta, and gamma diversities reduce to those based on ordinary Hill numbers.

Phylogenetic (Dis)similarity Measures

All the four classes of taxonomic (dis)similarity measures (left half in **Table 2**) can be readily generalized to their phylogenetic versions (right half in **Table 2**). In the generalization, the taxonomic beta diversity in the four types of transformations is replaced by the corresponding phylogenetic beta diversity. Therefore, all interpretations of the four classes of phylogenetic dissimilarity measures are parallel to their taxonomic version; the only difference in the interpretations is that species-equivalents in the taxonomic case is replaced by lineage-equivalents in the phylogenetic case (**Table 2**). For example, the phylogenetic Jaccard-type non-overlap measure $1 - \overline{U}_{qN}$ quantifies the effective proportion of non-shared lineages in the pooled assemblage, whereas the phylogenetic Sørensen-type non-overlap measure $1 - \overline{C}_{qN}$ quantifies the effective average proportion of non-shared lineages in an individual assemblage. The following special cases are high-lighted and are connected to the existing measures:

For the special case of $q = 0$, phylogenetic beta diversity becomes ${}^0PD = L_\gamma/L_\alpha$, where L_γ denotes the total branch length in the pooled assemblage, and L_α denotes the mean branch length per assemblage. Then the following branch-length-based similarity indices are obtained:

$\overline{S}_{0N} = \overline{U}_{0N} = \frac{L_\alpha/L_\gamma - 1/N}{1 - 1/N}$ (branch-length-based Jaccard similarity);

$\overline{V}_{0N} = \overline{C}_{0N} = \frac{N - L_\gamma/L_\alpha}{N-1}$ (branch-length-based Sørensen similarity).

For $N = 2$, $1 - \overline{S}_{02} = 1 - \overline{U}_{02}$ reduces to the UniFrac dissimilarity measure, and $\overline{V}_{02} = \overline{C}_{02}$ reduces to the PhyloSør similarity index.

When the element of the data matrix $[z_{ik}]$ represents the relative abundance of species i within the k-th assemblage, the phylogenetic Sørensen- and Jaccard-type non-overlap measures of $q = 1$ are identical, i.e., $1 - \overline{C}_{1N} = 1 - \overline{U}_{1N} = (H_{P,\gamma} - H_{P,\alpha})/T \log N$, where $H_{P,\gamma}$ and $H_{P,\alpha}$ denote gamma and alpha phylogenetic entropy. This measure represents the phylogenetic generalization of Horn's (1966) dissimilarity measure. For the special case of $q = 2$, the following phylogenetic Sørensen-type non-overlap measure shows how the additive beta based on the quadratic entropy (i.e., $Q_\gamma - Q_\alpha$) should be properly transformed to a legitimate dissimilarity index: $1 - \overline{C}_{2N} = (Q_\gamma - Q_\alpha)/[(1 - 1/N)(T - Q_\alpha)]$, where Q_γ and Q_α denote gamma and alpha quadratic entropy. Similarly, the corresponding transformation to obtain the phylogenetic Jaccard-type non-overlap measure is $1 - \overline{U}_{2N} = (Q_\gamma - Q_\alpha)/[(N - 1)(T - Q_\gamma)]$. These two measures represent phylogenetic generalizations of the Morisita-Horn dissimilarity and regional species non-overlap (Chiu *et al.*, 2014a) measures.

Regarding sampling effect, researchers in microbiology have found that branch-length based ($q = 0$) dissimilarity measures, such as the UniFrac index, are more sensitive to sampling effort/depth than abundance-sensitive measures (Lozupone *et al.*, 2011). Like the classic taxonomic richness-based dissimilarity indices, the observed UniFrac value generally exhibits positive bias, and more data are needed to detect more similar microbial communities. Based on sampling data, all asymptotic estimation and

standardization procedures for taxonomic beta diversity and species compositional (dis)similarity measures can be analogously generalized to their phylogenetic versions. Details are thus omitted; see a real data analysis below.

A Real-Data Example

Comparing Diversity of two Habitats in Forest Fragments

The tree datasets collected by Magnago and colleagues are used to illustrate the iNEXT standardization and its phylogenetic version (Magnago *et al.*, 2014). The tree species abundance data were collected between 2011 and 2012 from 24 transects (12 Edge transects and 12 Interior transects) in 11 rain-forest fragments in Espírito Santo State, in south-eastern Brazil. Each transect comprised ten 10 × 10 m plots at 20-m intervals. Each Edge transect was placed about 5 m inside the fragment and parallel to the forest edge, and each Interior transect was located at least 300 m from the nearest edge. Within each transect, every living tree with a diameter at breast height (DBH) > 4.8 cm and 1.3 m height was recorded. To compare diversity and compositional dissimilarity between the Edge habitat and the Interior habitat, data in each habitat were pooled from 12 transects.

The data considered here consist of 425 species represented by a total of 3868 individuals. Species abundance data and the phylogenetic tree spanned by the observed 425 species appear in Magnago *et al.* (2014) and Chao *et al.* (2017). The age of the root is about 400 Mys. There were 319 species among 1794 individuals in the data from the Edge habitat, and 356 species among 2074 individuals in the data from the Interior habitat. There were 250 species shared by the two habitats in the data. Although sample sizes for the two habitats are different, the sample coverage estimates for these two habitats are nearly equal (∼94% for both habitats). **Fig. 8** shows the two types (size- and coverage-based) of sampling curves for taxonomic diversity (Hill numbers) and mean phylogenetic diversity of orders $q = 0$, 1, and 2. Mean phylogenetic diversity (in units of lineages from the root) is used instead of phylogenetic diversity, because the former uses the same units as Hill numbers, allowing values to be directly compared across two dimensions of diversity.

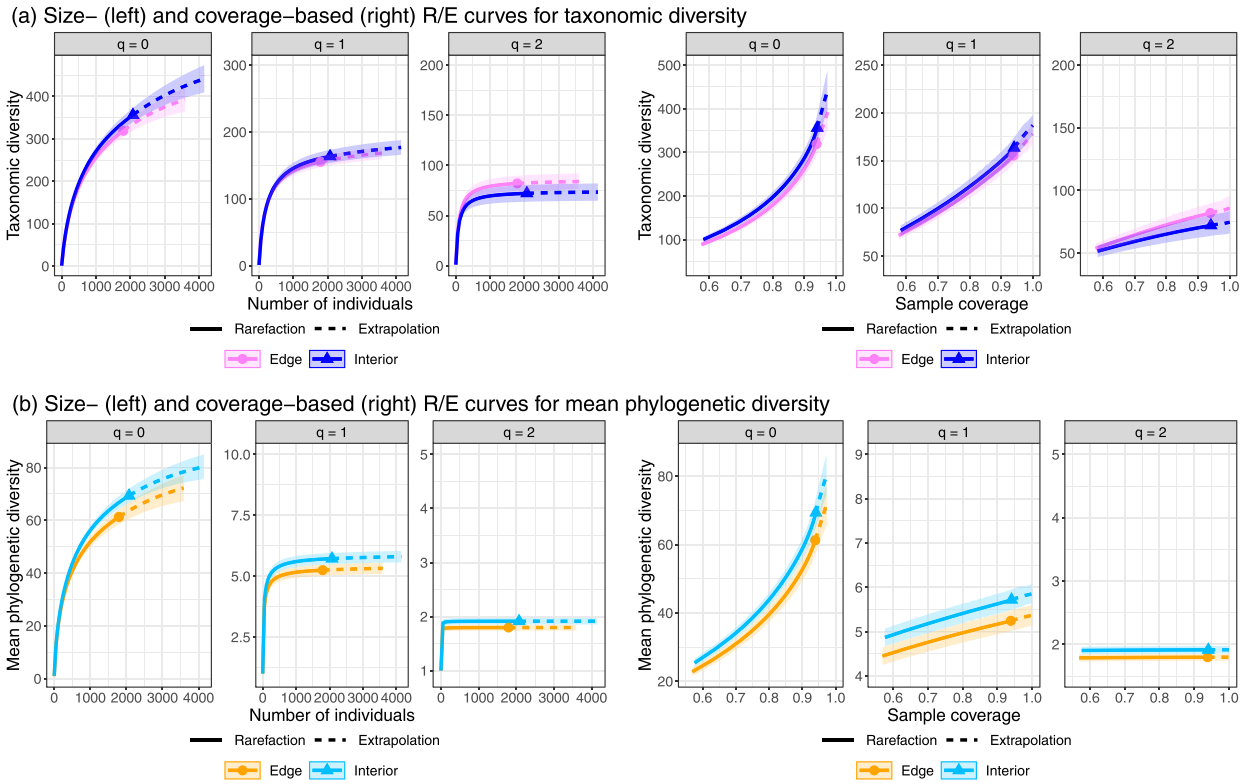

Fig. 8 Size- and coverage-based R/E (rarefaction/extrapolation) curves based on tree species data in two habitats (Edge and Interior) in rain forest fragments. Rarefaction (solid lines) and extrapolation (dashed lines) sampling curves are shown for (a) taxonomic diversity and (b) mean phylogenetic diversity of orders $q = 0$, 1, and 2. Observed (reference) sample points are denoted by solid dots (for Edge) and triangles (for Interior). All 95% confidence intervals (shaded areas) were based on a bootstrap method with 200 replications. In the size-based curves, the extrapolation extends up to double the reference sample size. In the coverage-based curves, the extrapolation extends up to 97% for $q = 0$, but up to complete coverage 100% for $q = 1$ and 2; see the text for details. Note that different scales are used in the *y*-axes. Data from Magnago, L. F. S., Edwards, D. P., Edwards, F. A., *et al.* (2014). Functional attributes change but functional richness is unchanged after fragmentation of Brazilian Atlantic forests. Journal of Ecology 102, 475–485.

As expected, the mean phylogenetic diversity (lower half in Fig. 8) is much lower than the corresponding taxonomic diversity (upper half) for any fixed sample size and any fixed sample coverage. (Notice the difference in the Y-axes scale in Fig. 8). This ordering can be explained by the fact that only tip nodes are used to compute taxonomic diversity and all tip nodes are assumed to be unrelated to each other. When evolutionary history is incorporated, species are revealed to be phylogenetically related, and thus taxonomic diversity represents the maximum value that mean phylogenetic diversity can attain.

Regardless of diversity type (taxonomic or phylogenetic), the left half of Fig. 8 reveals that, for each habitat, the size-based rarefaction and extrapolation sampling curves for diversity of orders $q = 1$ and $q = 2$ tend to stabilize, implying that the asymptotic diversity estimates for these two measures work satisfactorily to infer true diversities. However, none of the sampling curves for species/lineage richness ($q = 0$), extrapolated up to double reference size, reach a fixed level, suggesting that the current data do not contain sufficient information to accurately estimate true species/lineage richness within each habitat. In this case, fair comparison of species/lineage richness should be based on a standardized coverage value, as described below.

The sample coverage values for both reference samples are around 94%. When each sample is extrapolated to double the reference sample size, the coverage values for the two extrapolated samples are also very close (around 97%). Therefore, for $q = 0$, the coverage-based sampling curves are extrapolated up to 97%, which represents the highest coverage we can use for reliable comparison of standardized species/lineage richness estimates. In other words, the species/lineage richness of the remaining 3% of the assemblage is not estimable from the data, due to insufficient information for the 3% of individuals representing the rarest species. In this case, coverage-based sampling curves for $q = 0$ enable us to make sensible inferences and fair comparisons of diversity for any standardized assemblage fraction up to 97%. In contrast, for $q = 1$ and 2, data are sufficient to extrapolate the coverage-based sampling curves up to 100% (complete coverage).

The size- and coverage-based curves exhibit consistent diversity ordering between the two habitats. For taxonomic diversity, the ordering depends on the value of q: The Interior habitat is more diverse for $q = 0$ and 1, but the ordering is reversed for $q = 2$. This reversal signifies that the Edge habitat generally is more diverse than the Interior habitat in very abundant species, whereas it is less diverse in rare and abundant species, although the confidence bands do overlap. In contrast, when evolutionary history among species is considered, the Interior habitat is significantly more diverse than the Edge habitat, regardless of diversity order ($q = 0$, 1, and 2), as evidenced by the non-overlapping confidence intervals. That is, data are sufficient to show a significant deficit in the amount of evolutionary history for the species in the Edge habitat, regardless of species commonness and rarity. All the above conclusions for taxonomic and phylogenetic diversity comparisons are valid for coverage up to 97% for $q = 0$, and up to complete coverage for $q = 1$ and 2.

Assessing (Dis)similarity Between the two Habitats

To assess between-habitat compositional (dis)similarity, three estimation methods are considered: (a) Observed dissimilarity values based on the two reference samples. (b) Asymptotic dissimilarity estimates in which gamma and alpha were both estimated for the complete coverage value (100%). (c) Standardized dissimilarity estimates in which gamma and alpha were first rarefied to 97.36% (the highest matched coverage value to assess alpha diversity), then beta and the associated dissimilarity estimates at a coverage level of 97.36% were obtained. All numerical values presented in Table 3 were computed from the iNEXT.beta3D software (Chao *et al.*, 2023).

Table 3 shows the resulting dissimilarity values of orders $q = 0$, 1, and 2 based on two taxonomic non-overlap measures, $1 - C_{qN}$ (Sørensen-type) and $1 - U_{qN}$ (Jaccard-type); for $q = 1$, these two measures coincide. Table 3 also shows the two

Table 3 (a) Taxonomic and (b) phylogenetic observed, asymptotic, standardized (at 97.36% coverage) dissimilarity between the Edge and the Interior habitats for orders $q = 0$, 1, and 2, based on two taxonomic dissimilarity measures: $1 - C_{qN}$ (Sørensen-type species non-overlap) and $1 - U_{qN}$ (Jaccard-type species non-overlap) and two phylogenetic dissimilarity measures: $1 - \overline{C}_{qN}$ (Sørensen-type lineage non-overlap) and $1 - \overline{U}_{qN}$ (Jaccard-type lineage non-overlap); see Table 2 for the definitions and interpretations of each dissimilarity measure

Method	*Sørensen-type*			*Jaccard-type*		
	q = 0	*q = 1*	*q = 2*	*q = 0*	*q = 1*	*q = 2*
(a) Taxonomic dissimilarity						
Observed	0.259	0.249	0.440	0.412	0.249	0.282
Asymptotic	0.117	0.163	0.417	0.210	0.163	0.264
Standardized	0.036	0.141	0.412	0.070	0.141	0.259
(b) Phylogenetic dissimilarity						
Observed	0.208	0.065	0.014	0.345	0.065	0.007
Asymptotic	0.079	0.050	0.013	0.146	0.050	0.007
Standardized	0.046	0.046	0.013	0.089	0.046	0.007

Data from Magnago, L. F. S., Edwards, D. P., Edwards, F. A. *et al.* (2014). Functional attributes change but functional richness is unchanged after fragmentation of Brazilian Atlantic forests. Journal of Ecology 102, 475 – 485.

corresponding phylogenetic non-overlap measures, $1 - \overline{C}_{qN}$ (Sørensen-type) and $1 - \overline{U}_{qN}$ (Jaccard-type). The other two taxonomic and phylogenetic turnover measures generally lead to consistent patterns, and thus are omitted.

First notice that for $q = 2$, the three estimation methods for each fixed measure yield comparable taxonomic, and nearly identical phylogenetic dissimilarity values. These results suggest that dissimilarity measures of $q = 2$ tend to be resistant to undersampling, because the most abundant species/lineages are always present in samples. In contrast, for measures of $q = 0$ and $q = 1$, the observed dissimilarity values are typically higher than the asymptotic values. It is most likely the case that the observed dissimilarity values of $q = 0$ and $q = 1$ over-estimate the true values, which is consistent with general findings in the literature.

As with the diversity inference discussed earlier, the asymptotic dissimilarity values of $q = 1$ and 2 can be used to infer the true dissimilarity. These two abundance-sensitive measures ($q = 1$ and 2) are mainly dominated by nodes/branches with relatively high abundances and long branch lengths; such nodes appear near the root (where the differentiation is near zero) and most of those influential nodes are shared. Thus, phylogenetic dissimilarity is generally lower than the corresponding taxonomic dissimilarity. For example, based on the Sørensen-type taxonomic non-overlap measures, the asymptotic dissimilarity values for $q = 1$ and 2 are $1 - C_{1N} = 16.3\%$ and $1 - C_{2N} = 41.7\%$, which means, effectively, that the average percentage of abundant/common species that are unique to each habit is about 16.3%, and about 41.7% for very abundant species. The corresponding phylogenetic non-overlap measures are much lower, with $1 - \overline{C}_{1N} = 5\%$ and $1 - \overline{C}_{2N} = 1.3\%$. These values can be interpreted to mean that the effective average percentage of abundant lineages that are unique to each habitat is about 5%, and about 1.3% for very abundant lineages. Similar interpretations can be made for the Jaccard-type non-overlap measures.

However, for $q = 0$, data are not sufficient to accurately infer the true dissimilarity of $q = 0$. Accurate dissimilarity can be assessed only up to a coverage value of 97%. Based on the standardized Sørensen-type dissimilarity values at a coverage level of 97%, taxonomic dissimilarity values are $1 - C_{0N} = 3.6\%$, $1 - C_{1N} = 14.1\%$ and $1 - C_{2N} = 41.2\%$. The stronger effects for $q = 1$ and 2 imply that the species compositional dissimilarity was principally attributable to abundant and very abundant species, i.e., mainly linked to a shift in the species abundance distribution, rather than to a change in species richness. This conclusion is also supported by the standardized taxonomic Jaccard-type dissimilarity measure. In contrast, both the standardized phylogenetic Sørensen- and Jaccard-type dissimilarity measures decline with order q, revealing that phylogenetic dissimilarity was principally attributable to lineage richness, i.e., linked to a shift in the amount of evolutionary history, rather than node abundance distribution. These taxonomic and phylogenetic (dis)similarity analyses provide meaningful comparisons and shed light on different contributions of the abundant/rare species or lineages to biodiversity difference between the Edge and Interior habitats in rainforest fragments.

Conclusions

This article introduces a unified framework based on Hill numbers of order q (Eqs. (2a) and (2b)) and their phylogenetic generalizations (Eqs. (5) and (6)) to quantify taxonomic and phylogenetic diversity in an assemblage. To assess among-assemblage taxonomic and phylogenetic differentiation across space and/or time, the framework also features diversity decomposition; **Table 1** shows pertinent taxonomic alpha, beta, and gamma formulas and their phylogenetic versions. The resulting beta diversity can be monotonically transformed to several interpretable (dis)similarity measures (**Table 2**). Based on incomplete sampling data, two procedures are recommended to control for sampling effects: (1) asymptotic diversity estimation (including species-richness estimation) to reduce undersampling bias, and (2) a non-asymptotic method, implemented in iNEXT, via size- and coverage-based rarefaction and extrapolation to standardize sampling effort or sampling completeness. These analyses provide meaningful and robust comparison of diversity and the associated (dis)similarity measures. **Fig. 8** and **Table 3** show how these standardization measures can be applied to real-world examples to extract ecological meaning from complex biodiversity data.

References

Alberdi, A. and Gilbert, M. T. P. (2019) A guide to the application of Hill numbers to DNA-based diversity analyses. Molecular Ecology Resources 19, 804–817.

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

Bryant, J. A., Lamanna, C. and Morlon, H. *et al.* (2008) Microbes on mountainsides: Contrasting elevational patterns of bacterial and plant diversity. Proceedings of the National Academy of Sciences, USA 105, 11505–11511.

Burnham, K. P. and Overton, W. S. (1979) Robust estimation of population size when capture probabilities vary among animals. Ecology 60, 927–936.

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

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

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

Chao, A. and Chiu, C.-H. (2016) Bridging two major approaches (the variance framework and diversity decomposition) to beta diversity and related similarity and differentiation measures. Methods in Ecology and Evolution 7, 919–928.

Chao, A., Chiu, C. H. and Colwell, R. K. *et al.* (2017) Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory. Ecology 98, 2914–2929.

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

Chao, A. and Colwell, R. K. (2017) Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling. Statistics and Operation Research Transactions (SORT) 41, 3–54.
Chao, A., Colwell, R. K., Lin, C.-W. and Gotelli, N. J. (2009) Sufficient sampling for asymptotic minimum species richness estimators. Ecology 90, 1125–1133.
Chao, A., Gotelli, N. G. and Hsieh, T. C. *et al.* (2014) Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species biodiversity studies. Ecological Monographs 84, 45–67.
Chao, A., Henderson, P. A. and Chiu, C.-H. *et al.* (2021) Measuring temporal change in alpha diversity: a framework integrating taxonomic, phylogenetic and functional diversity and the iNEXT.3D standardization. Methods in Ecology and Evolution 12, 1926–1940.
Chao, A. and Jost, L. (2012) Coverage-based rarefaction: standardizing samples by completeness rather than by size. Ecology 93, 2533–2547.
Chao, A. and Jost, L. (2015) Estimating diversity and entropy profiles via discovery rates of new species. Methods in Ecology and Evolution 6, 873–882.
Chao, A., Jost, L., Chiang, S. C., Jiang, Y.-H. and Chazdon, R. (2008) A two-stage probabilistic approach to multiple-community similarity indices. Biometrics 64, 1178–1186.
Chao, A., Ma, K.H., Hsieh, T.C. and Chiu, C.H. (2015). Online Program SpadeR (Species-richness prediction and diversity estimation in R). Available at: https://chao.shinyapps.io/SpadeR/.
Chao, A. and Shen, T. J. (2004) Nonparametric prediction in species sampling. Journal of Agricultural, Biological, and Environmental Statistics 9, 253–269.
Chao, A., Thorn, S., and Chiu, C.-H. *et al.* (2023) Rarefaction and extrapolation with beta diversity under a framework of Hill numbers: the iNEXT.beta3D standardization. To appear in Ecological Monographs. https://doi.org/10.1002/ecm.1588.
Chiu, C.-H., Jost, L. and Chao, A. (2014a) 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. and Chao, A. (2014b) An improved non-parametric lower bound of species richness via a modified Good-Turing frequency formula. Biometrics 70, 671–682.
Coleman, B. D., Mares, M. A., Willig, M. R. and Hsieh, Y.-H. (1982) Randomness, area, and species richness. Ecology 63, 1121–1133.
Colwell, R. K., Chao, A. and Gotelli, N. J. *et al.* (2012) Models and estimators linking individual-based and sample-based rarefaction, extrapolation, and comparison of assemblages. Journal of Plant Ecology 5, 3–21.
Colwell, R. K. and Coddington, J. A. (1994) Estimating terrestrial biodiversity through extrapolation. Philosophical Transactions of the Royal Society B 345, 101–118.
Colwell, R. K., Mao, C. X. and Chang, J. (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. Ecology 85, 2717–2727.
Cormack, R. M. (1989) Log-linear models for capture-recapture. Biometrics 45, 395–413.
Faith, D. P. (1992) Conservation evaluation and phylogenetic diversity. Biological Conservation 61, 1–10.
Ferrier, S., Manion, G., Elith, J. and Richardson, K. (2007) Using generalized dissimilarity modelling to analyse and predict patterns of beta diversity in regional biodiversity assessment. Diversity and Distribution 13, 252–264.
Fisher, R. A., Corbet, A. S. and Williams, C. B. (1943) The relation between the number of species and the number of individuals in a random sample of an animal population. Journal of Animal Ecology 12, 42–58.
Good, I. J. (1953) The population frequencies of species and the estimation of population parameters. Biometrika 40, 237–264.
Gotelli, N. J. and Colwell, R. K. (2001) Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. Ecology Letters 4, 379–391.
Gotelli, N. J. and Colwell, R. K. (2011) Estimating species richness. In: Magurran, A., McGill, B. (eds.) Biological diversity: Frontiers in measurement and assessment. Oxford: Oxford University Press, pp. 39–54.
Hill, M. O. (1973) Diversity and evenness: A unifying notation and its consequences. Ecology 54, 427–431.
Horn, H. S. (1966) Measurement of "overlap" in comparative ecological studies. American Naturalist 100, 419–424.
Hsieh, T. C. and Chao, A. (2017) Rarefaction and extrapolation: making fair comparison of abundance-sensitive phylogenetic diversity among multiple assemblages. Systematic Biology 66, 100–111.
Hsieh, T. C., Ma, K. H. and Chao, A. (2016) iNEXT: An R package for rarefaction and extrapolation of species diversity (Hill numbers). Methods in Ecology and Evolution 7, 1451–1456.
Hurlbert, S. H. (1971) The nonconcept of species diversity: A critique and alternative parameters. Ecology 52, 577–586.
Jost, L. (2007) Partitioning diversity into independent alpha and beta components. Ecology 88, 2427–2439.
Jost, L., Chao, A. and Chazdon, R. L. (2011) Compositional similarity and beta diversity. In: Magurran, A., McGill, B. (eds.) Biological diversity: Frontiers in measurement and assessment. Oxford: Oxford University Press, pp. 66–84.
Legendre, P. and De Cáceres, M. (2013) Beta diversity as the variance of community data: Dissimilarity coefficients and partitioning. Ecology Letters 16, 951–963.
Lozupone, C. and Knight, R. (2005) UniFrac: A new phylogenetic method for comparing microbial communities. Applied and Environmental Microbiology 71, 8228–8235.
Lozupone, C., Lladser, M. E., Knights, D., Stombaugh, J. and Knight, R. (2011) UniFrac: An effective distance metric for microbial community comparison. The ISME Journal 5, 169–172.
MacArthur, R. H. (1965) Patterns of species diversity. Biological Reviews 40, 510–533.
Magnago, L. F. S., Edwards, D. P. and Edwards, F. A. *et al.* (2014) Functional attributes change but functional richness is unchanged after fragmentation of Brazilian Atlantic forests. Journal of Ecology 102, 475–485.
Magurran, A.E. (2004). Measuring biological diversity. Oxford: Blackwell.
Morisita, M. (1959) Measuring of interspecific association and similarity between communities. Memoires of the Faculty of Science, Kyushu University, Series E (Biolology) 3, 65–80.
Peet, R. K. (1974) The measurement of species diversity. Annual Review of Ecology and Systematics 5, 285–307.
Pielou, E.C. (1975). Ecological diversity. New York: John Wiley.
Rao, C. R. (1982) Diversity and dissimilarity coefficients: A unified approach. Theoretical Population Biology 21, 24–43.
Sanders, H. (1968) Marine benthic diversity: A comparative study. American Naturalist 102, 243–282.
Shen, T. J., Chao, A. and Lin, C. F. (2003) Predicting the number of new species in further taxonomic sampling. Ecology 84, 798–804.
Shinozaki, K. (1963). Notes on the species-area curve. In: Proceedings of the 10th Annual Meeting of the Ecological Society of Japan (Abstract), 5.
Simpson, E. H. (1949) Measurement of diversity. Nature 163, 688.
Tokeshi, M. (1999) Species coexistence: Ecological and evolutionary perspectives. Oxford: Blackwell.
Vane-Wright, R. I., Humphries, C. J. and Williams, P. M. (1991) What to protect: systematics and the agony of choice. Biological Conservation 55, 235–254.
Whittaker, R. H. (1960) Vegetation of the Siskiyou mountains, Oregon and California. Ecological Monographs 30, 279–338.
Whittaker, R. H. (1972) Evolution and measurement of species diversity. Taxon 12, 213–251.

Further Reading

Chao, A. and Colwell, R. K. (2022) Biodiversity: concepts, dimensions, and measures. pp 25-46 in "The Ecological and Societal Consequences of Biodiversity Loss" (coordinated by Loreau, M., Hector, A., and Isbell, F.) John Wiley, New York.
Colwell, R. K. and Chao, A. (2022) Measuring and comparing class diversity in archaeological assemblages: A brief guide to the history and state-of-the-art in diversity statistics. pp. 263–294 in "Defining and Measuring Diversity in Archaeology: Another Step Toward an Evolutionary Synthesis of Culture" (Eren, M. I., and B. Buchanan, editors). Berghahn, New York.
Ellison, A. M., Gotelli, N. J., Farnsworth, E. J. and Alpert, G. D. (2012). A field guide to the ants of New England. New Haven, CT: Yale University Press.

Relevant Websites

https://chao.shinyapps.io/iNEXT_beta3D/.
Interpolation and Extrapolation with beta diversity.
https://chao.shinyapps.io/iNEXTOnline/.
Interpolation/Extrapolation (iNEXT).
https://chao.shinyapps.io/SpadeR/.
Species-richness prediction and diversity estimation with R (SpadeR).