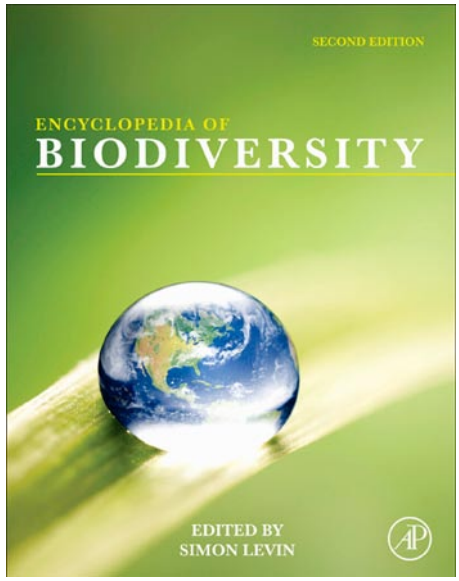

Gotelli Nicholas J., and Chao Anne (2013) Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data. In: Levin S.A. (ed.) Encyclopedia of Biodiversity, second edition, Volume 5, pp. 195-211. Waltham, MA: Academic Press.

Measuring and Estimating Species Richness, Species Diversity, and Biotic Similarity from Sampling Data

Nicholas J Gotelli, University of Vermont, Burlington, VT, USA
Anne Chao, National Tsing Hua University, Hsin-Chu, Taiwan

Glossary

Biotic similarity A measure of the degree to which two or more samples or assemblages are similar in species composition. Familiar biotic similarity indices include Sørensen's, Jaccard's, Horn's, and Morisita's indices.

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.

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

Phylogenetic diversity Adjusted diversity measures that take into account the degree of relatedness 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 by drawing random subsets of individuals (or samples) in order to standardize the comparison of biological diversity on the basis of a common number of individuals or samples.

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, and transect lines). The incidence or presence of each species is recorded for each sampling unit.

Species accumulation curve A curve of rising biodiversity in which the x-axis is the number of sampling units (individuals or samples) from an assemblage and the y-axis is the observed species richness. The species accumulation curve rises monotonically to an asymptotic maximum number of species.

Species diversity A measure of diversity that incorporates both the number of species in an assemblage and some measure 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 samples collected. Species richness is a diversity of order 0 (which means it is completely insensitive to species abundances).

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 that is 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" are 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 the distinct species. These definitions are generic, 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 (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 (understory forest herbs and filter-feeding molluscs). 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 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 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 species increases can reflect deliberate and accidental introductions or range shifts driven by climate change.

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 (**Figure 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" 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 receive equal weighting, and (3) diversity is measured in appropriate units (individuals, biomass, and percentage cover are most commonly used).

Phylogenetic, Taxonomic, 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 evenness 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 **Figure 2** for another example).

The concept of traditional diversity can therefore be extended to consider differences among species. All else being equal, an assemblage of phylogenetically or functionally divergent 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 taxonomic classification (referred to as taxonomic diversity) or phylogeny (referred to as phylogenetic

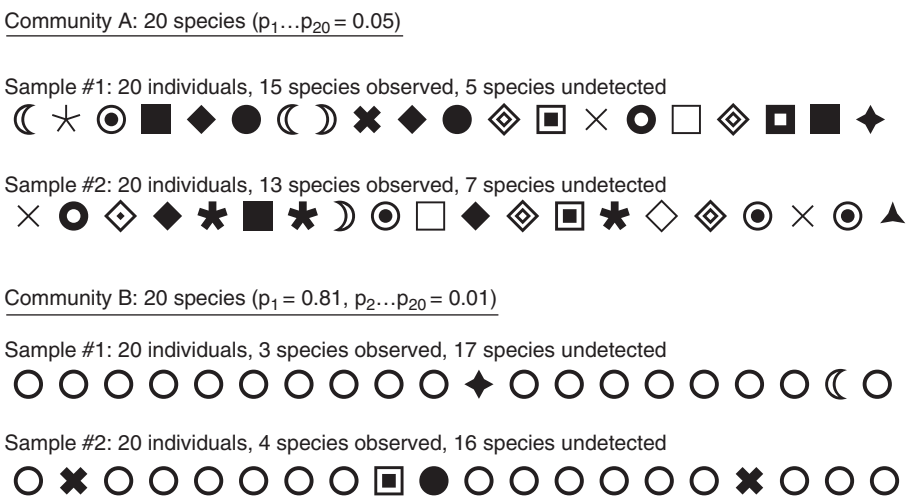

Figure 1 Species richness sampling in a hypothetical walk through the woods. Each different symbol represents one of 20 distinct species, and each row contains 20 characters, representing the first 20 individual trees that might be encountered in a random sample. Community A is maximally even, with each 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.

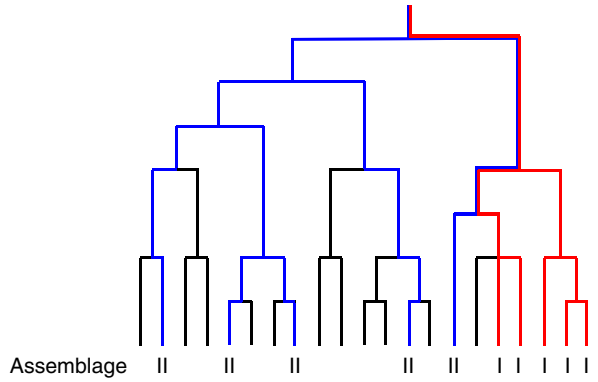

Figure 2 Phylogenetic diversity in species composition. The branching diagram is a hypothetical phylogenetic tree. The ancestor of the entire assemblage is the "root" at the top, with time progressing toward the branch tips at the bottom. Each node (branching point) represents a speciation or divergence event, and the 21 branch tips illustrate the 21 extant species. Extinct species or lineages are not illustrated. The five species in Assemblage I represent an assemblage of five closely related species (they all share a quite recent common ancestor). The five species in Assemblage II 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.

diversity (PD)) or indirectly, based on their function (referred to as functional diversity). These metrics relax the second assumption discussed in the section Species Richness and Traditional Species Diversity Metrics (all species are "equally different" from one another) by weighting each species by a measure of its taxonomic classification, phylogeny, or function.

Biotic 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, both of which have 10 trees species, each species contributing 10% to the abundance of individual trees within the 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 is 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 losses 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, and biotic similarity must be estimated from samples taken from the assemblage. If the sample relative abundances are used directly in the formulas for traditional diversity and similarity measures, the maximum likelihood estimator (MLE) of the true diversity or similarity measure is obtained. However, the MLEs of most species diversity measures are biased when sample sizes are small. 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 usually 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 the 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 the 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 will reflect sampling uncertainty.

The Organization of Biodiversity Sampling Data

This article introduces a common set of notation for describing biodiversity data (Colwell *et al.*, 2012). Consider an assemblage consisting of N^* total individuals, each belonging to one of S distinct species. Species i has N_i individuals, so that $\sum_{i=1}^{S} N_i = N^*$. The relative frequency p_i of species i is N_i/N^*, so that $\sum_{i=1}^{S} p_i = 1$. Note here that N^*, S, N_i, and p_i represent the "true" underlying abundance, species richness, and relative frequencies of species. These quantities are unknowns, but 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 N^* total individuals. In the reference sample, a total of S_{obs} species are observed, with X_i individuals observed for species i, so that $\sum_{i=1}^{S} X_i = n$ (only species with $X_i > 0$ contributes 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 N^* individuals and S species, but were 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 replicate *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 presence or absence of each species is recorded, but abundances or counts of the species present are not needed. 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.

The 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 (those that are each detected in only one sample) and Q_2 represents the number of "duplicate" species (those that are each detected in exactly two samples). The zero frequency Q_0 denotes the number of species among the S species in the 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 (**Figure 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, as the common species in the assemblage are detected relatively quickly. The curve continues to rise as

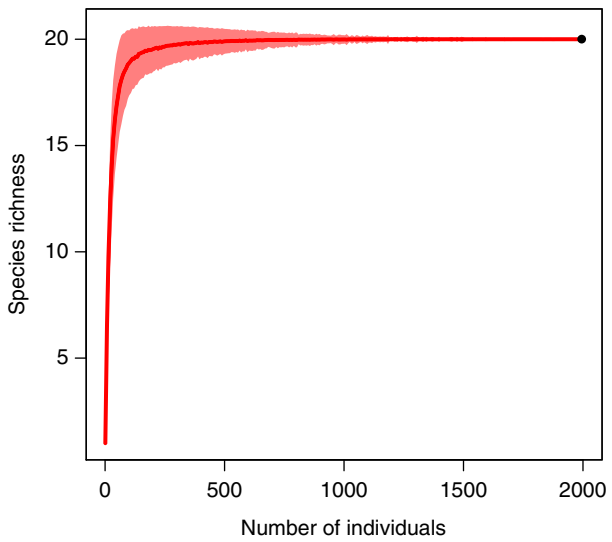

Figure 3 Species accumulation curve. The curve was generated by assuming an assemblage of 20 species whose relative abundances were created from a broken stick distribution (Tokeshi, 1999). The x-axis is the number of individuals sampled and the y-axis is the number of species observed. The species accumulation 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.

more individuals are sampled, but the slope becomes shallower because progressively more sampling is required to detect the rare species. As long as the sampling area is sufficiently homogeneous, 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, a similar 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 samples, they must be standardized to a common number of individuals, for abundance samples (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 (**Figure 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

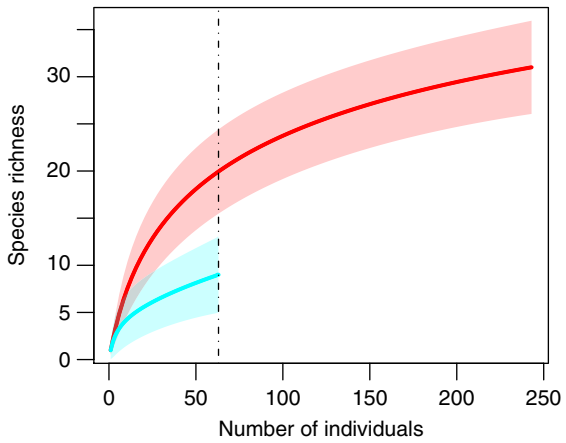

Figure 4 Standardized comparison of species richness for two individual-based rarefaction curves. The data represent summary counts of carabid beetles that were pitfall-trapped from a set of young pine plantations (<20 years old; upper curve) and a set of old pine plantations (20–60 years old; lower curve). The solid lines are the rarefaction curves, calculated from eqn [2], and the shaded polygons are the 95% confidence intervals, calculated from the unconditional variance eqn [5]. 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.

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

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

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

$$\tilde{S}_{\text{ind}}(m) = S_{\text{obs}} - \sum_{X_i>0} \left[\binom{n - X_i}{m} \Big/ \binom{n}{m} \right] \qquad [2]$$

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

$$\tilde{S}_{\text{sample}}(r) = S_{\text{obs}} - \sum_{Y_i>0} \left[\binom{R - Y_i}{r} \Big/ \binom{R}{r} \right] \qquad [3]$$

where $r<R$ is the number of sampling units in the rarefied reference sample. The statistical model for rarefaction is sampling without replacement from the reference sample.

If the area of each of the sample plots 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 and is a function of the total area A of the archipelago (or the summed areas of all the sample plots) (Coleman *et al.*, 1982):

$$\tilde{S}_{\text{area}}(a) = S_{\text{obs}} - \sum_{X_i>0} \left(1 - \frac{a}{A}\right)^{X_i} \qquad [4]$$

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 generally should widen as the reference sample size is reached. This unconditional variance for abundance data is calculated as follows (Colwell *et al.*, 2012):

$$\sigma^2_{\text{ind}}(m) = \sum_{k=1}^{n} (1 - \alpha_{km})^2 f_k - [\tilde{S}_{\text{ind}}(m)]^2 / S_{\text{est}} \qquad [5]$$

where S_{est} denotes an estimated species richness (such as Chao1, described in the section Species Richness Estimation) and $\alpha_{km} = \binom{n-k}{m} \Big/ \binom{n}{m}$ for $k \le n - m$, $\alpha_{km} = 0$ otherwise. The corresponding unconditional variance for incidence data is (Colwell *et al.*, 2004)

$$\sigma^2_{\text{sample}}(r) = \sum_{k=1}^{R} (1 - \beta_{kr})^2 Q_k - [\tilde{S}_{\text{sample}}(r)]^2 / S_{\text{est}} \qquad [6]$$

where S_{est} denotes a sample-based estimated species richness (such as Chao2, described in the section Species Richness

Estimation) and $\beta_{kr} = \binom{R-k}{r} \Big/ \binom{R}{r}$ for $k \leq R-r$, $\beta_{kr}=0$ otherwise. These variances allow for the calculation of 95% confidence intervals on expected species richness for any abundance level smaller than the observed sample (**Figure 4**).

Because all rarefaction curves converge at small sample sizes toward 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), random spatial arrangement of individuals, and random, independent sampling of individuals (or larger sampling units for incidence-based methods). If the spatial distribution of individuals is intraspecifically clumped in space, abundance-based rarefaction will overestimate species richness, but this problem can be effectively countered by increasing the spatial grain of sampling or using incidence-based methods. 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. However, calculation and comparison of complete rarefaction curves and their extrapolation, with unconditional variances, help 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. *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) to the species accumulation data, and the asymptote can be estimated as one of the parameters of this kind of model. The chief problem is that this 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*, 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. It is often not clear that existing statistical models fit empirical data sets very 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):

$$\hat{S}_{\text{Chao1}} = \begin{cases} S_{\text{obs}} + f_1^2/(2f_2) & \text{if } f_2 > 0 \\ S_{\text{obs}} + f_1(f_1-1)/2 & \text{if } f_2 = 0 \end{cases} \quad [7]$$

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

$$\hat{var}(\hat{S}_{\text{Chao1}}) = f_2\left[\frac{1}{2}\left(\frac{f_1}{f_2}\right)^2 + \left(\frac{f_1}{f_2}\right)^3 + \frac{1}{4}\left(\frac{f_1}{f_2}\right)^4\right] \quad [8]$$

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

$$\hat{S}_{\text{Chao2}} = \begin{cases} S_{\text{obs}} + [(R-1)/R]Q_1^2/(2Q_2) & \text{if } Q_2 > 0 \\ S_{\text{obs}} + [(R-1)/R]Q_1(Q_1-1)/2 & \text{if } Q_2 = 0 \end{cases} \quad [9]$$

with a variance estimator of (if $Q_2>0$)

$$\hat{var}(\hat{S}_{\text{Chao2}}) = Q_2\left[\frac{A}{2}\left(\frac{Q_1}{Q_2}\right)^2 + A^2\left(\frac{Q_1}{Q_2}\right)^3 + \frac{1}{4}A^2\left(\frac{Q_1}{Q_2}\right)^4\right] \quad [10]$$

where $A=(R-1)/R$. When $f_2=0$ in the Chao1 estimator or $Q_2=0$ in the Chao 2 estimator, the variance formulas in

eqns [9] and [10] need modification; the modified variances are available in Chao and Shen (2010).

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 $\leq \kappa$) 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 $S_{abun} = \sum_{i>\kappa} f_i$ and the number of observed species in the rare species group be $S_{rare} = \sum_{i=1}^{\kappa} f_i$. Define $n_{rare} = \sum_{i=1}^{\kappa} i f_i$ and the coverage estimate $\hat{C}_{rare} = 1 - f_1/n_{rare}$. Coverage is the estimated proportion of the total number of N^* individuals in the assemblage that is represented by the species recorded in the sample. It is a reliable measure of the degree of sample completeness. The *Abundance-based Coverage Estimator* (*ACE*) is (Chao, 2005)

$$\hat{S}_{ACE} = S_{abun} + \frac{S_{rare}}{\hat{C}_{rare}} + \frac{f_1}{\hat{C}_{rare}}\hat{\gamma}^2_{rare} \qquad [11]$$

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

$$\hat{\gamma}^2_{rare} = \max\left\{\frac{S_{rare}}{\hat{C}_{rare}}\frac{\sum_{i=1}^{\kappa} i(i-1)f_i}{(\sum_{i=1}^{\kappa} i f_i)(\sum_{i=1}^{\kappa} i f_i - 1)} - 1, 0\right\} \qquad [12]$$

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

For incidence data, there is a corresponding *Incidence-based Coverage Estimator* (*ICE*). As with ACE, first a cut-off point κ is selected 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 $S_{freq} = \sum_{i>\kappa} Q_i$ and the number of species in the infrequent group by $S_{infreq} = \sum_{i=1}^{\kappa} Q_i$. The estimated sample coverage for the infrequent group is $\hat{C}_{infreq} = 1 - Q_1/\sum_{i=1}^{\kappa} i Q_i$. Let the number of sampling units that include at least one infrequent species be R_{infreq}. Then ICE is expressed as

$$\hat{S}_{ICE} = S_{freq} + \frac{S_{infreq}}{\hat{C}_{infreq}} + \frac{Q_1}{\hat{C}_{infreq}}\hat{\gamma}^2_{infreq} \qquad [13]$$

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

$$\hat{\gamma}^2_{infreq} = \max\left\{\frac{S_{infreq}}{\hat{C}_{infreq}}\frac{R_{infreq}}{(R_{infreq}-1)} \times \frac{\sum_{i=1}^{\kappa} i(i-1)Q_i}{(\sum_{i=1}^{\kappa} i Q_i)(\sum_{i=1}^{\kappa} i Q_i - 1)} - 1, 0\right\} \qquad [14]$$

In addition to Chao1, Chao2, ACE, and ICE, the *jackknife* method 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 subdata by successively deleting j individuals from the data. The first-order jackknife turns out to be

$$\hat{S}_{jk1} = S_{obs} + \frac{n-1}{n} f_1 \approx S_{obs} + f_1 \qquad [15a]$$

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

$$\hat{S}_{jk2} = S_{obs} + \frac{2n-3}{n} f_1 - \frac{(n-2)^2}{n(n-1)} f_2 \approx S_{obs} + 2f_1 - f_2 \qquad [15b]$$

Higher-order jackknife estimators are available, although they give increasingly less weight to the more common species frequencies.

For replicated incidence data, the first-order jackknife for R samples is

$$\hat{S}_{jk1} = S_{obs} + \frac{R-1}{R} Q_1 \qquad [16a]$$

and the second-order jackknife is

$$\hat{S}_{jk2} = S_{obs} + \frac{2R-3}{R} Q_1 - \frac{(R-2)^2}{R(R-1)} Q_2 \qquad [16b]$$

All jackknife estimators can be expressed as linear combinations of frequency counts, and thus approximate variances and confidence intervals can be directly obtained.

Extrapolating Species Richness

Based on a reference sample of n individuals, the extrapolation or prediction problem is to estimate the expected number of species $S_{ind}(n+m^*)$ 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:

$$\tilde{S}_{ind}(n+m^*) = S_{obs} + \hat{f}_0\left[1 - \left(1 - \frac{f_1}{n\hat{f}_0}\right)^{m^*}\right] \approx S_{obs} + \hat{f}_0\left[1 - \exp\left(-\frac{m^* f_1}{n\hat{f}_0}\right)\right] \qquad [17]$$

where $\hat{f}_0$ is an estimator for f_0 (the number of undetected species). It is suggested that $\hat{f}_0$ can be obtained by using either the Chao1 estimator ($\hat{f}_0 = \hat{S}_{Chao1} - S_{obs}$) or the ACE estimator ($\hat{f}_0 = \hat{S}_{ACE} - S_{obs}$).

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

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

$$\tilde{S}_{\text{area}}(A+a^*) = S_{\text{obs}} + \hat{f}_0\left[1-\exp\left(-\frac{a^*}{A}\frac{f_1}{\hat{f}_0}\right)\right] \quad [18]$$

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

For sample-based data with R sampling units comprising the reference sample, Chao *et al.* (2009, Appendix A) 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:

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

$$\approx S_{\text{obs}} + \hat{Q}_0\left[1-\exp\left(\frac{-r^*Q_1}{Q_1+R\hat{Q}_0}\right)\right] \quad [19]$$

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 $(\hat{Q}_0=\hat{S}_{\text{ICE}}-S_{\text{obs}})$.

For each of the 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 (**Figure 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).

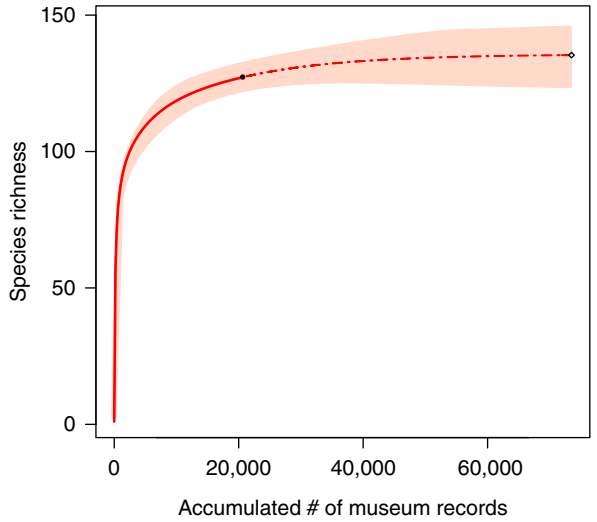

Figure 5 A smoothed rarefaction and extrapolation curve. The *x*-axis is the number of individual, geo-referenced, dated ant specimens in New England, and the *y*-axis is the observed number of species. The total collection (the reference sample, filled circle) included 127 species and 20,225 individual 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. 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 AM, Gotelli NJ, Farnsworth EJ, and Alpert GD (2012) *A Field Guide to the Ants of New England*. New Haven, CT: Yale University Press.

Species 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 measure 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).

For example, 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. However, as with the estimation of asymptotic richness, these methods often do not perform well unless the "true" species abundance distribution is known, which is never the case (Colwell and Coddington, 1994; Chao, 2005).

Nonparametric methods make no assumptions about the mathematical form of the underlying species abundance distribution, and they have been widely used not only in ecology but also in information science, economics, genetics, and linguistics (see Jost, 2007; Jost *et al.*, 2011; Chao and Jost, in press; Tuomisto, this volume for reviews). The most popular of these measures is the *Shannon entropy*,

$$H_{\text{Sh}} = -\sum_{i=1}^{S} p_i \log p_i \quad [20]$$

where S is the number of species in the assemblage and the ith species relative abundance is p_i. Shannon entropy quantifies the uncertainty in the species identity of a randomly chosen individual in the assemblage. Another measure that has been widely used in economics and genetics, as well as in ecology, is the Gini–Simpson index,

$$H_{\text{GS}} = 1-\sum_{i=1}^{S} p_i^2 \quad [21]$$

which measures the probability that two randomly chosen individuals (selected with replacement) belong to two different species. The measure $1 - H_{GS} = \sum_{i=1}^{S} p_i^2$ is referred to as the Simpson index. With an adjustment for N^*, 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:

$$\text{PIE} = [N^*/(N^* - 1)]H_{GS} \qquad [22]$$

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. These popular complexity 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 $\exp(H_{Sh})$, and the Gini–Simpson index can be transformed to $1/(1 - H_{GS}) = 1/\sum_{i=1}^{S} p_i^2$, 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 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 Gini–Simpson measures do not obey the "replication principle." However 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. Different Hill numbers ${}^qD$ are defined by their "order" q as (Hill, 1973)

$${}^qD = \left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)} \qquad [23a]$$

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

$${}^1D = \lim_{q \to 1} {}^qD = \exp\left(-\sum_{i=1}^{S} p_i \log p_i\right) = \exp(H_{sh}) \qquad [23b]$$

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 (no "discounting" for uneven abundances), and ${}^0D$ equals species richness. When $q = 1$, the Hill number ${}^1D$ is the exponential form of Shannon entropy, which weighs species in proportion to their frequency and can be roughly interpreted as the number of "typical species" in the assemblage (Chao *et al.*, 2010; Chao and Jost, in press). When $q = 2$, ${}^2D$ equals $1/(1 - H_{GS})$, which heavily weights the most common species in the assemblage; the contribution from rare species is severely discounted. The measure ${}^2D$ can be roughly 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 plots surveyed) than the most popular Hill numbers ($q = 0$, 1, 2). Hill numbers with negative exponents can also be calculated, but they place so much weight on rare species they have poor sampling properties.

Thus, the measure of diversity using Hill numbers can potentially depend on the order q that is chosen. However, because all Hill numbers need not be integers, and all have common units of species richness, they can be portrayed on a single graph as a function of q. This "diversity profile" of effective species richness versus q portrays all of the information about species abundance distribution of an assemblage (**Figure 6**). 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.

Estimation of Hill Numbers

All of the Hill numbers (including species richness) as well as the untransformed Gini–Simpson index and Shannon entropy are sensitive to the number of individuals and samples collected. The sample-size dependence diminishes as q increases because the higher-order Hill numbers are more heavily weighted by frequencies of common species, and the estimates of those frequencies are not very sensitive to sample size. In contrast, with increasing numbers of individuals or samples collected, rare species continue to be added to the sample, making richness and other low Hill numbers more sample size dependent.

The MLE for the Gini–Simpson index, $\hat{H}_{GS,MLE} = 1 - \sum_{i=1}^{S} (X_i/n)^2$, is biased downward, and the bias in some cases can be substantial. The minimum variance unbiased estimator (MVUE) of the Gini–Simpson index has the following relationship to its MLE:

$$\begin{aligned}\hat{H}_{GS,MVUE} &= 1 - \sum_{i=1}^{S} [X_i(X_i - 1)]/[n(n-1)] \\ &= [n/(n-1)]\hat{H}_{GS,MLE}\end{aligned} \qquad [24a]$$

This MVUE is equivalent to the estimator of *PIE* and is relatively invariant to sample size. Thus, a nearly unbiased estimator for Hill number of order 2 is

$${}^2\hat{D} = 1/\sum_{i=1}^{S} [X_i(X_i - 1)]/[n(n-1)] \qquad [24b]$$

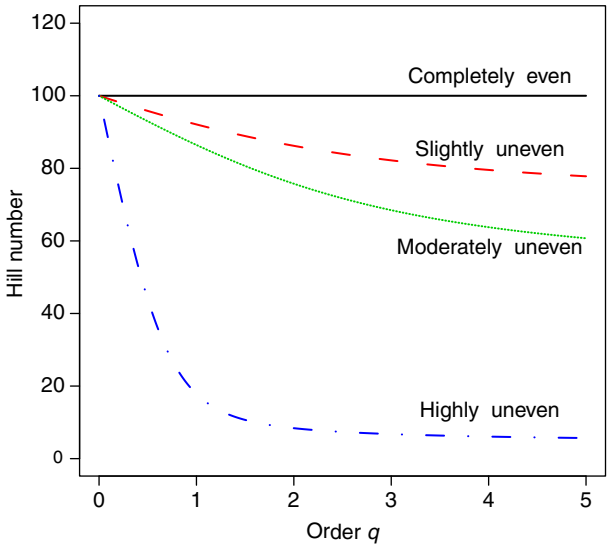

Figure 6 Diversity profile for assemblages of differing evenness. The *x*-axis is the order q in the Hill number (eqn [23a]), and is illustrated for values of q from 0 to 5. The *y*-axis is the calculated Hill number (the equivalent number of equally abundant species). Each of the four assemblages has 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 $\{50 \times 7, 50 \times 3\}$); (3) moderately uneven assemblage (green dotted line): $\{22 \times 10, 28 \times 5, 40 \times 3, 10 \times 2\}$; (4) highly uneven assemblage (blue dash–dot line): $\{1 \times 120, 1 \times 80, 1 \times 70, 1 \times 50, 3 \times 20, 3 \times 10, 90 \times 1\}$. 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.

For an integer $q \geq 2$, a similar derivation leads to a nearly unbiased estimator for ${}^{q}D$.

$$ {}^{q}\hat{D} = \left\{ \sum_{i=1}^{S} [X_i(X_i-1)\cdots(X_i-q+1)] \Big/ [n(n-1)\cdots(n-q+1)] \right\}^{1/(1-q)} \quad [24c] $$

These estimators for $q \geq 2$ are almost independent of sample size, because all the higher-order Hill numbers are mainly dominated by the number of very abundant species. The estimated diversity profile curve is thus generally slowly varying for $q \geq 2$.

The estimation of entropy has been well studied in information science, physics, and statistics. Unfortunately, an unbiased estimator for Shannon entropy does not exist for any fixed sample size of n. As noted earlier, using X_i/n as a simple estimator of the true p_i value yields the MLE of entropy, which is negatively biased. An estimator of Shannon entropy with low bias is the following Horvitz–Thompson-type estimator (Chao and Shen, 2003):

$$ \hat{H}_{\mathrm{Sh}} = -\sum_{i=1}^{S} \frac{\tilde{p}_i \log(\tilde{p}_i)}{1-(1-\tilde{p}_i)^n} \quad [25a] $$

where $\tilde{p}_i = (X_i/n)(1 - f_1/n)$ is an estimator of the true p_i. Only the detected species contribute to the summation because $\tilde{p}_i = 0$ for any undetected species. The denominator $1-(1-\tilde{p}_i)^n$ is the estimated probability that the ith species is detected in the sample, and the inverse of this probability is used as a weight for the ith species. Thus, the larger the probability of detection, the smaller the weight in the Horvitz-Thompson estimator. The weights adjust the estimator to compensate for missing species. For $q=1$, a low-bias estimator of the Hill number is

$$ {}^{1}\hat{D} = \exp\left(-\sum_{i=1}^{S} \frac{\tilde{p}_i \log(\tilde{p}_i)}{1-(1-\tilde{p}_i)^n} \right) \quad [25b] $$

In summary, the statistical properties of Hill numbers depend on the order q. Equation [24b] is a nearly unbiased estimator of diversity for $q=2$, and eqn [25b] is a low-bias estimator for $q=1$. As discussed in the section Species richness estimation, total species richness (${}^{0}D=S$) is much more difficult to estimate because it is very sensitive to rare species that are often undetected, even in relatively large samples. Several nonparametric species richness estimators that can be used for estimating ${}^{0}D$ are provided in the section Nonparametric Asymptotic Species Richness Estimators. Then an estimated diversity profile can be constructed by plotting $\{{}^{q}\hat{D}; q = 0,1,2,3,\ldots\}$ with respect to q, based on estimators given in eqns [24b], [24c], and [25b]. The variance of each estimator in the profile can be approximated by a standard asymptotic method, and a 95% confidence interval can thus be constructed as the estimator ± 1.96 s.e. for each value of q, if the sample size is sufficiently large.

Taxonomic and PD

To quantify taxonomic or PD, species are placed on a branching tree (a cladogram) that describes their evolutionary relationships (**Figure 2**). The base of the tree represents the ancestral taxon, the branching forks (nodes) represent speciation or divergence events, the branch tips represent the contemporary species (not all of which may be represented in any particular assemblage), and time is measured in the vertical axis, increasing from the base of the tree to the branch tips. (For paleontological applications, the tips may be extinct lineages.) 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.

This article distinguishes two types of phylogenetic trees: ultrametric and nonultrametric trees. A tree is called ultrametric if all branch tips are the same distance from the basal node. For example, if the branch lengths are proportional to divergence time, the tree is ultrametric. 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. In contrast, if the branch lengths are proportional to the number of base-pair changes in a given gene, or some other measure of genetic or morphological change, some branch tips may be farther in absolute time from the basal node than other branch tips, and such trees are nonultrametric.

Pielou (1975) was the first to notice that the concept of diversity could be broadened to consider differences among species. The earliest taxonomic diversity measure 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 (1992) *PD*, 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*, species abundances are not considered.

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

$$Q_{\text{Rao}} = \sum_{i,j} d_{ij} p_i p_j \qquad [26a]$$

where d_{ij} denotes the phylogenetic distance between species i and j, and p_i and p_j denote the relative abundance of species i and j. This index measures the average phylogenetic 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_{Rao} 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):

$$H_P = -\sum_i L_i a_i \log a_i \qquad [26b]$$

where the summation is over all branches, L_i is the length of branch i, and a_i denotes the summed abundance of all species descended from branch i. The notation a (abundance) here is not the same as that (for area) in eqn [4].

The replication principle can be generalized to phylogenetic or functional diversity: When N completely distinct trees (no shared nodes during the fixed time interval of interest) with equal diversities are combined, the diversity of the combined tree is N times the diversity of any individual tree. Simple examples can show that neither Q_{Rao} nor H_p satisfies this replication principle. As with traditional diversity measures, Q_{Rao} and H_p can be transformed into measures that do obey the replication principle. In an ultrametric tree with tree height T^* (the time interval between the tree base and tips), the transformed measures are, respectively, $\exp(H_p/T^*)$ and $1/[1-(Q_{\text{Rao}}/T^*)]$.

For ultrametric trees, in addition to the order q, a time parameter T is required to generalize Hill numbers to measure PD in an interval from $-T$ time steps to the present time. The generalized phylogenetic metric is a time-averaged measure of lineage diversity (Hill numbers) at any moment t over the time interval $[-T, 0]$. The lineage diversity ${}^qD(t)$ at any moment t is measured by taking a "cross-section" and finding the lineages that are intersected (**Figure 7**). The relative abundance of each lineage is defined as the sum of the relative abundances of all of the descendants of that lineage in the present-day assemblage. Thus, ${}^qD(t)$ can be quantified by a Hill number. The average of these Hill numbers ${}^qD(t)$ over the interval $[-T, 0]$ gives the *mean PD* of order q over T time steps (Chao *et al.*, 2010):

$$ {}^q\overline{D}(T) = \left\{ \sum_{i \in B_T} \frac{L_i}{T} a_i^q \right\}^{1/(1-q)} \qquad [27]$$

where $\boldsymbol{B}_T$ denotes the set of all branches in the time interval $[-T, 0]$, L_i is the length (duration) of branch i in the set $\boldsymbol{B}_T$, and a_i is the total abundance of extant species descending from branch i; see **Figure 7** for a hypothetic ultrametric tree.

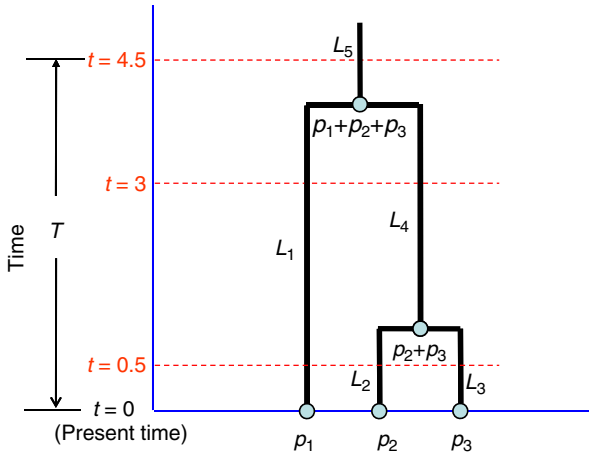

Figure 7 Calculation of mean phylogenetic diversity and branch diversity. In this hypothetical rooted phylogenetic tree, the ancestor to the assemblage is depicted at the top, and there are three extant species living at the present, depicted at the bottom, with relative abundances $(p_1, p_2, p_3) = (0.2, 0.3, 0.5)$. The tree is ultrametric, so the total branch length from the ancestor to any descendant species in the present is the same. To evaluate the phylogenetic diversity at the time $T = 5$ time steps in the past, we first create the set $\boldsymbol{B}_T$ which includes five branches with lengths $(L_1, L_2, L_3, L_4, L_5) = (4, 1, 1, 3, 1)$ and the corresponding abundances $(a_1, a_2, a_3, a_4, a_5) = (p_1, p_2, p_3, p_2+p_3, p_1+p_2+p_3)$. We next measure the lineage diversity ${}^qD(t)$ at any time steps $0 < t < T$. We use three different "sampling times" as examples; these three sampling times correspond to the three distinct assemblages that would be represented by diversity sampling at three points in the past. For the first sampling time, $t=0.5$, lineage diversity ${}^qD(t)$ is measured as the Hill numbers for three lineages (species) with relative abundances (p_1, p_2, p_3); for the second sampling time $t=3$, lineage diversity ${}^qD(t)$ is measured as the Hill numbers for two lineages (species) with relative abundances (p_1, p_2+p_3); for the third sampling time $t=4.5$, lineage diversity ${}^qD(t)$ is measured as the Hill numbers for only one lineage (species) with relative abundance $p_1+p_2+p_3=1$. The average of these Hill numbers ${}^qD(t)$ over the interval $[-T, 0]$ gives the *mean phylogenetic diversity* ${}^q\overline{D}(T)$ of order q over T time steps. The branch diversity is ${}^qPD(T) = T \times {}^q\overline{D}(T)$. These two diversities can be calculated for any fixed $T \geq 0$; their pattern as a function of T is plotted in **Figure 8**. (For example, if T is changed to the tree height $T^*=4$ (distance between tree base and tips), then the branch set includes four branches $(L_1, L_2, L_3, L_4)=(4, 1, 1, 3)$ with the corresponding abundances $(a_1, a_2, a_3, a_4)=(p_1, p_2, p_3, p_2+p_3)$, and the lineage diversity is averaged over $[-4, 0]$ to obtain ${}^q\overline{D}(T^*)$ and ${}^qPD(T^*) = T^* \times {}^q\overline{D}(T^*)$.)

This measure ${}^{q}\overline{D}(T)$ gives the mean effective number of maximally distinct lineages (or species) T time steps in the past. The diversity of a tree with ${}^{q}\overline{D}(T)=z$ in the time period $[-T, 0]$ is the same as the diversity of an assemblage consisting of z equally abundant and maximally distinct species with all branch lengths T.

The branch diversity or phylogenetic diversity ${}^{q}\mathrm{PD}(T)$ of order q through T time steps before present is defined as the product of ${}^{q}\overline{D}(T)$ and T . The measure ${}^{q}\mathrm{PD}(T)$ given below quantifies "the total effective number of lineage-lengths or lineage-time steps" (Chao *et al.*, 2010)

$$ {}^{q}\mathrm{PD}(T) = T \times {}^{q}\overline{D}(T) = \left\{ \sum_{i \in B_T} L_i \left(\frac{a_i}{T} \right)^{q} \right\}^{1/(1-q)} \qquad [28] $$

If $q=0$ and $T=T^{*}$ (tree height), then ${}^{0}\mathrm{PD}(T)$ reduces to Faith's PD. It also reduces to CD in a taxonomic tree if the branching of each Linnaean taxonomic category is assigned a time step of unit length. A PD profile can be constructed by plotting both ${}^{q}\mathrm{PD}(T)$ and ${}^{q}\overline{D}(T)$ as a function of T for $q=0$, 1, and 2. It is also informative to construct another diversity profile by plotting ${}^{q}\mathrm{PD}(T)$ and ${}^{q}\overline{D}(T)$ as a function of order q for some selected values of temporal perspective T. See **Figure 8** for a numerical example. In most applications, ecologists are interested in the case $T=T^{*}$ (tree height) or the divergence time between the species group of interest and its nearest outgroup. The divergence time of the most recent common ancestor of all extant taxa is another useful comparison.

For nonultrametric trees, the time parameter T is generalized to $\overline{T}$, where $\overline{T}=\sum_{i \in B_{\overline{T}}} L_i a_i$ represents the abundance-weighted mean base change per species and $\boldsymbol{B}_{\overline{T}}$ denote the set of branches connecting all focal species. The diversity of a nonultrametric tree with mean evolutionary change $\overline{T}$ is the same as that of an ultrametric tree with a time step $\overline{T}$.

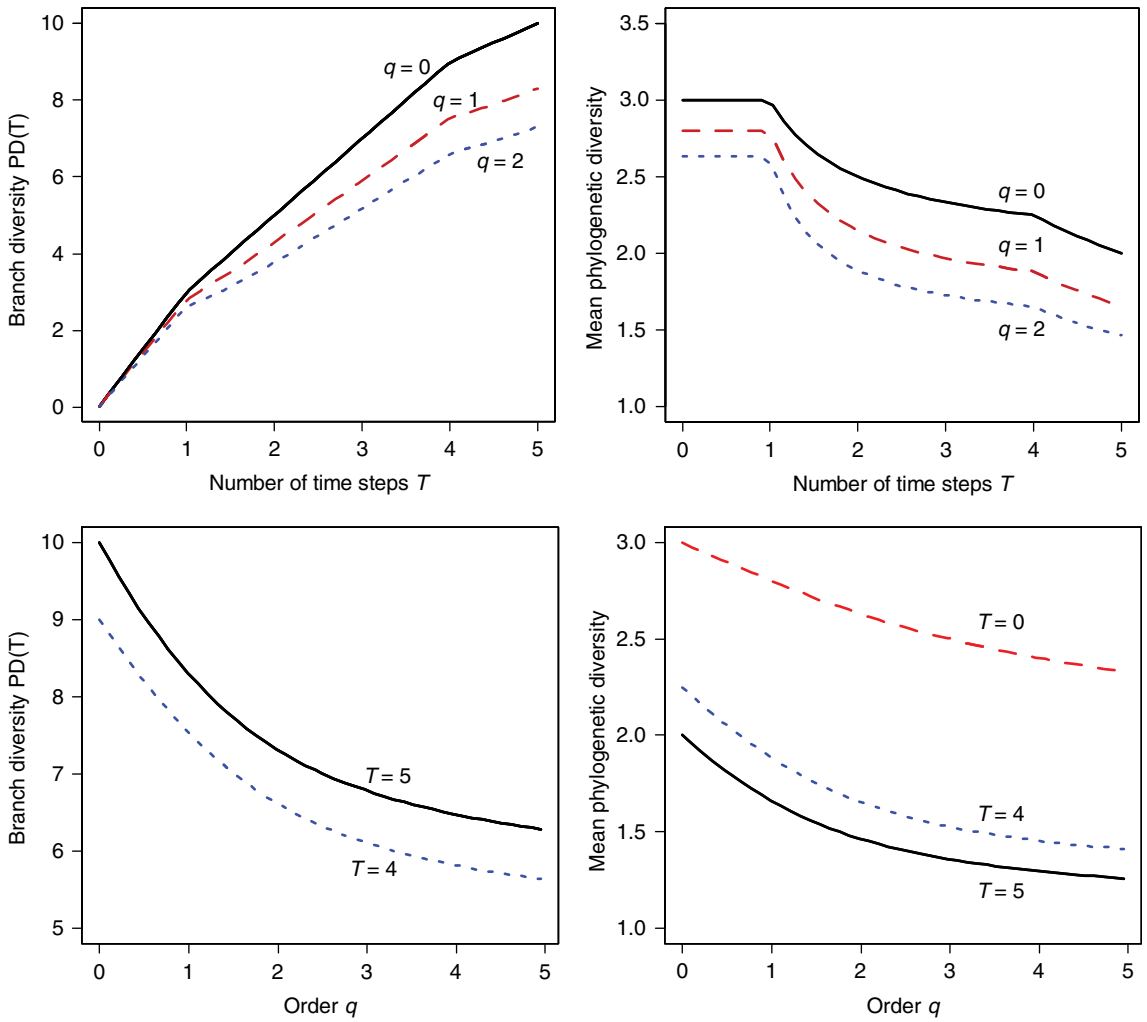

Figure 8 Branch diversity profile and mean phylogenetic diversity profile. (a) Branch diversity profile of ${}^{q}\mathrm{PD}(T)$ (upper left panel) and mean phylogenetic diversity (upper right panel) as a function of the number of time steps (T) in the past ($0<T<5$) for $q=0$, 1, and 2, based on the structure of the phylogenetic tree in **Figure 7**, assuming $(p_1, p_2, p_3)=(0.2, 0.3, 0.5)$ and $(L_1, L_2, L_3, L_4, L_5)=(4, 1, 1, 3, 1)$. (b) Branch diversity profile of ${}^{q}\mathrm{PD}(T)$ (lower left panel) as a function of order q for $T=4$ (tree height) and $T=5$ time steps, and mean phylogenetic diversity profile of (lower right panel) as a function of order q for $T=0$, 4, and 5, assuming $(p_1, p_2, p_3)=(0.2, 0.3, 0.5)$ and $(L_1, L_2, L_3, L_4, L_5)=(4, 1, 1, 3, 1)$. The branch diversity for $T=0$ is 0 as all branch lengths are 0. The mean phylogenetic diversity for $T=0$ are the traditional Hill numbers for $(p_1, p_2, p_3)=(0.2, 0.3, 0.5)$.

Therefore, the diversity formula for a nonultrametric tree is obtained by replacing T in ${}^{q}\overline{D}(T)$ and ${}^{q}\mathrm{PD}(T)$ with $\overline{T}$. Equation [27] can also describe taxonomic diversity, if the phylogenetic tree is a Linnaean tree with L levels (ranks), and each branch is assigned unit length. It also describes functional diversity, if a dendrogram can be constructed from a trait-based distance matrix using a clustering scheme (Petchey and Gaston, 2002). Thus, Hill numbers can be effectively generalized to incorporate taxonomy, phylogeny, and function and provide a unified framework for measuring biodiversity (Chao and Jost, in press).

Estimation of phylogenetic and functional diversity from small samples has not been well studied. As with the estimation of simple Hill numbers, phylogenetic diversity ${}^{q}\overline{D}(T)$ and ${}^{q}\mathrm{PD}(T)$ can be accurately estimated only for $q=2$. Like the Gini–Simpson index, the MVUE of Rao's quadratic entropy exists under a multinomial model:

$$\hat{Q}_{\mathrm{MVUE}} = \sum_{i,j} d_{ij} X_i X_j / [n(n-1)] \qquad [29]$$

Thus, ${}^{2}\overline{D}(T)$ and ${}^{2}\mathrm{PD}(T)$ can be estimated by a nearly unbiased measure using this estimator based on the transformation ${}^{2}\overline{D}(T) = 1/[1-(Q_{\mathrm{Rao}}/T)]$. With sufficient sampling, these estimators of PD of order $q=2$ are almost independent of sample size. Further research is needed for the development of accurate estimators of PD measures with $q=1$ and 0.

Biotic Similarity

Incidence-Based Similarity Indices

The earliest published incidence-based measure of relative compositional similarity is the classic Jaccard index from 1900. A number of incidence-based similarity measures have been proposed since then (see Jost *et al.*, 2011, for a review). The Jaccard index and the Sørensen index (proposed in 1948) 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$=S_{12}/(S_1+S_2-S_{12})$ and the Sørensen similarity index$=2S_{12}/(S_1+S_2)$. A rearrangement of the Sorensen index$=1/[0.5(S_{12}/S_1)^{-1}+0.5(S_{12}/S_2)^{-1}]$ reveals that it is the harmonic mean of two proportions: S_{12}/S_1 (the proportion of the species in the first assemblage that are shared with the second) and S_{12}/S_2 (the proportion of the species in the second assemblage that are shared with the first). The Jaccard index compares the number of shared species to the total number of species in the combined assemblages, whereas the Sørensen index compares the number of shared species to the mean number of species in a single assemblage. The Jaccard index is thus a comparison based on total diversity, whereas the Sørensen index is a comparison based on local diversity.

When one assemblage is much richer than the other, both Sørensen and Jaccard indices become very small. Although the low similarity value reflects the true difference between the two assemblages, in some applications it can be more informative to normalize a similarity measure so that maximum overlap$=1.0$. Lennon *et al.* (2001) proposed such a modification to the Sørensen index, and it takes the form $S_{12}/\min(S_1, S_2)$; see Jost *et al.* (2011) for details and comparisons.

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, the pairwise similarities calculated from data tend to be correlated and are not independent. Most importantly, pairwise similarities cannot fully characterize multiple-assemblage similarity when some species are shared across two, three, or more assemblages (Chao *et al.*, 2008). It is easy to construct numerical examples in which all pairwise similarities are identical in two sets of assemblages, but the global similarities for the two sets are different.

The two-assemblage incidence-based Jaccard and Sørensen indices have been extended to multiple assemblages. Assume that there are N assemblages and there are S_j species in the jth assemblage and S species in the combined assemblage. Let $\overline{S}$ denote the average number of species per assemblage. The multiple-assemblage Jaccard similarity index$=(\overline{S}/S-1/N)/(1-1/N)$. The multiple-assemblage Sørensen similarity index$=(N-S/\overline{S})/(N-1)$. When $N=2$, these two measures reduce to their classical two-assemblage measures. These two measures are decreasing functions of Whittaker's beta diversity for species richness, which is $S/\overline{S}$. When N assemblages are identical, beta diversity ($q=0$) is $S/\overline{S}=1$, and thus both Jaccard and Sørensen similarity indices$=1$. When N assemblages are completely distinct (no shared species), beta diversity ($q=0$) is $S/\overline{S}=N$, and thus both Jaccard and Sørensen similarity indices$=0$.

These incidence-based similarity indices are widely used in ecology and biogeography because of their simplicity and easy interpretation. In most ecological studies, these indices are estimated from observed richness in sample data. The resulting estimates are generally biased downward, and the bias increases when sample sizes are small or species richness is large. They could become biased upward when shared species are common and endemic species are very rare (Chao *et al.*, 2005, p. 149). The classic pairwise Jaccard and Sorensen similarity indices calculated from sample data generally underestimate the true similarity mainly because they do not account for shared species at both sites that were not detected. One strategy could be to use asymptotic species richness estimators (*see* Species Richness Estimation) to estimate species richness in each assemblage and also to estimate species richness in the combined assemblage, and then substitute the estimated values into the similarity formulas. However, this strategy inevitably inflates the variance and often renders the resulting estimate useless. A major statistical concern is that, based on incidence data alone, bias correction and measurements of variances are impossible. Consequently, the interpretation of any incidence-based index based on sample values or estimated values becomes difficult or misleading for comparing two (or more) highly diverse assemblages based on limited data. Only with abundance data can one correct for undersampling bias, as explained in the next section.

Classical incidence-based similarity indices treat abundant and rare species 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.

Abundance-Based Similarity Indices

Assume that in the combined assemblages, there are S species. Denote the relative abundance vector for the S species in the jth assemblage by $(p_{1j},p_{2j},\ldots,p_{sj})$, some of them may be 0. Thus, for N assemblages, there are N sets of abundances $\{(p_{1j},p_{2j},\ldots,p_{sj});\ j=1,2,\ldots,N\}$. A sample of n_j individuals is taken from the jth assemblage and there are N sets of sample frequencies $\{(X_{1j},X_{2j},\ldots,X_{sj});\ j=1,2,\ldots,N\}$.

For two-assemblage cases, one of the most popular abundance-based similarity metric is the Horn overlap measure (Horn, 1966), which is based on Shannon's entropy:

$$S_H = \frac{1}{\log 2}\sum_{i=1}^{S}\left[\frac{p_{i1}}{2}\log\left(1+\frac{p_{i2}}{p_{i1}}\right)+\frac{p_{i2}}{2}\log\left(1+\frac{p_{i1}}{p_{i2}}\right)\right] \quad [30]$$

Another popular overlap measure is the Morisita–Horn similarity measure (Morisita, 1959), based on the Simpson index:

$$S_{MH} = \frac{\sum_{i=1}^{S} p_{i1}p_{i2}}{\left[\sum_{i=1}^{S} p_{i1}^2+\sum_{i=1}^{S} p_{i2}^2\right]/2} = 1-\frac{\sum_{i=1}^{S}(p_{i1}-p_{i2})^2}{\sum_{i=1}^{S} p_{i1}^2+\sum_{i=1}^{S} p_{i2}^2} \quad [31]$$

Since each of the indices [30] and [31] equals unity if and only if $p_{i1}=p_{i2}$ for all i, these two indices match relative abundances on a species-by-species basis. When the two assemblages are equally diverse and consist entirely of equally common species, the Morisita–Horn index, the Horn index, and the Sørensen index are all equal, and all of these indices give the proportion of shared species in an assemblage.

The first expression in eqn [31] for the Morisita-Horn index has an important probabilistic interpretation. If one individual is selected randomly from each assemblage, then the probability that the two selected individuals belong to *the same shared* species is $\sum p_{i1}p_{i2}$, the numerator in eqn [31]. The denominator in eqn [31] represents a normalizing constant, which is the average of two such probabilities for two individuals drawn from the same assemblages. In this probabilistic interpretation, the abundant species will contribute the most to the probability that two randomly selected individuals belong to the *same* species. As a result, in a hyperdiverse assemblage, the index is dominated by a few abundant species and the relatively rare species (even if there are many of them) have little effect. The index is therefore likely to be resistant to undersampling, because the influential abundant species are always present in samples.

The following MLE of the Morisita–Horn index is always in the range [0, 1], but it has been shown that this MLE systematically underestimates the true similarity.

$$\tilde{S}_{MH,\ MLE} = \frac{2\sum_{i=1}^{S}(X_{i1}/n_1)(X_{i2}/n_2)}{\sum_{i=1}^{S}(X_{i1}/n_1)^2+\sum_{i=1}^{S}(X_{i2}/n_2)^2} \quad [32]$$

A better estimator that is nearly unbiased has the following form, although it may exceed the theoretical maximum value of 1:

$$\tilde{S}_{MH} = \frac{2\sum_{i=1}^{S}(X_{i1}/n_1)(X_{i2}/n_2)}{\sum_{i=1}^{S}(X_{i1}(X_{i1}-1)/[n_1(n_1-1)])+\sum_{i=1}^{S}(X_{i2}(X_{i2}-1)/[n_2(n_2-1)])} \quad [33]$$

It is statistically infeasible to derive analytic or asymptotic variance formulas for these two estimators and the other similarity estimators in this section. A bootstrap method, suggested in Chao *et al.* (2008), is a simple and direct data-resampling method to obtain approximate estimates of variances and confidence intervals especially for complex estimators (Efron and Tibshirani, 1993). This method has found wide applications in various disciplines.

To the extent that ecological processes are often most strongly influenced by abundant species, the Morisita–Horn measure is useful when looking for functional differences between ecosystems. However, when rare species are important, as in many conservation applications, the Horn overlap measure would be more useful. But the MLE of the Horn overlap measure exhibits moderate bias due to undersampling. Until now, only the two-sample jackknife technique has been used to remove part of bias (Jost *et al.*, 2011). More research is required to find a reliable bias-reduced estimator.

For assessing similarity among more than two assemblages, a general multiple-assemblage, abundance-based overlap measure C_{qN} (Chao *et al.*, 2008) is

$$C_{qN} = \frac{[1/(N^q-N)]\sum_{i=1}^{S}\left[(p_{i1}+p_{i2}+\cdots+p_{iN})^q-(p_{i1}^q+p_{i2}^q+\cdots+p_{iN}^q)\right]}{(1/N)\sum_{i=1}^{S}(p_{i1}^q+p_{i2}^q+\cdots+p_{iN}^q)} \quad [34]$$

As with the Hill numbers, here q is a parameter that determines the measure's sensitivity to species' relative abundances, and N is the number of assemblages. The C_{qN} measure includes, as special cases, the classic two-assemblage Sørensen index ($q=0$, $N=2$), the Horn overlap index ($q=1$, $N=2$), the Morisita–Horn similarity index ($q=2$, $N=2$), and their multiple-assemblage generalizations ($N>2$) as follows:

For $q=0$, C_{0N} is the multiple-assemblage Sørensen similarity index:

$$C_{0N} = (N-S/\overline{S})/(N-1) \quad [35a]$$

For $q=1$, C_{1N} is the multiple-assemblage Horn overlap index:

$$C_{1N} = \frac{1}{\log N}\sum_{i=1}^{S}\sum_{j=1}^{N}\left[\frac{p_{ij}}{N}\log\left(1+\frac{\sum_{k\neq j}p_{ik}}{p_{ij}}\right)\right] \quad [35b]$$

For $q=2$, C_{2N} is the multiple-assemblage Morisita–Horn similarity index:

$$C_{2N} = \frac{2\sum_{i=1}^{S}\sum_{j<k}p_{ij}p_{ik}}{(N-1)\sum_{i=1}^{S}\sum_{j=1}^{N}p_{ij}^2} \quad [35c]$$

Jost (2007) was the first to develop a rigorous mathematical formulation of alpha and beta diversities based on Hill numbers of order q. He derived the multiplicative beta diversity ${}^qD_\beta$, which quantifies the effective number of completely distinct assemblages; see also Jost *et al.* (2011), Chao and Jost (in press), and Tuomisto (this volume) for reviews. For

equally weighted assemblages, beta diversity ${}^{q}D_{\beta}$ ranges between a minimum of 1 (when all assemblages are identical) and a maximum of N, the number of assemblages in each region (when all assemblages are completely distinct; i.e., there are no shared species). For example, a set of completely distinct sites in a region of three sites attains the maximum value of 3, whereas another set of completely distinct sites in a region of 10 sites attains the maximum value of 10. Because the maximum depends on the number of assemblages in the region, beta diversities usually cannot be compared directly among multiple regions. Instead, beta diversity should be compared with sample-based rarefaction to a common number of samples or to a common degree of completeness of samples in each region. However, beta diversity can be transformed to the C_{qN} measure in the range [0, 1] by the following nonlinear transform for N equally weighted assemblages:

$$C_{qN} = [(1/{}^{q}D_{\beta})^{q-1} - (1/N)^{q-1}]/[1 - (1/N)^{q-1}] \qquad [36]$$

The transformed measure C_{qN} is unity (when all assemblages are identical) and 0 (when all assemblages are completely distinct).

This nonlinear transformation ensures that C_{qN} preserves an essential property of an overlap index: The transformed index C_{qN} gives the true *overlap* A/S for *all* orders of q if N assemblages each have S equally common species, exactly A species are shared by all of them, and the remaining species are not shared by any assemblages. No linear transformation of beta diversity can achieve this property. See **Figure 9**, Case I, for an example. The measure C_{qN} thus quantifies the *effective average overlap per assemblage*, i.e., average *percentage of overlapped species* (species that are shared by all assemblages, as defined in the above sense) in an assemblage. Because this measure is in a sense of "average" overlap, it can be compared across regions with different number of assemblages. See Chao *et al.* (2008) for more details on the interpretation of C_{qN} based on a simple reference set of assemblages.

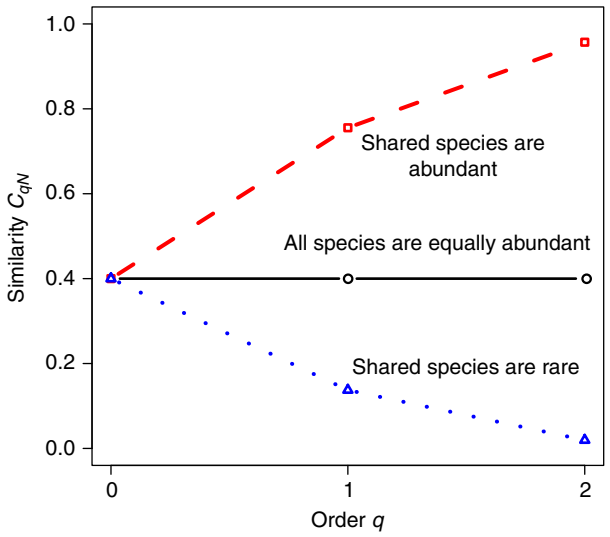

Figure 9 The two-assemblage similarity profile C_{qN} for $q=0$, 1, 2. *Case* I: In each of Assemblages 1 and 2, there are 20 equally abundant species, and eight species are shared. Thus, for all orders of q, the similarity measure C_{qN} equals the percentage of overlap in each assemblage. That is, $C_{qN}=8/20=40\%$ for all q (*black solid line*).
Case II: In each of Assemblages 1 and 2, there are 20 species, and the relative abundances in each assemblage are $p_i \propto K/i$ for $i=1, 2, \ldots, 20$, where K is a normalizing constant such that the total relative abundances is 1. Assume that the shared species are the most abundant eight species. The similarity measure C_{qN} increases as order q is increased (*red dashed line*).
Case III: Same as in Case II, but the shared species are the rarest eight species. The similarity measure C_{qN} decreases as order q is increased (*blue dotted line*).

Just as diversity profiles are used to characterize traditional diversity (**Figure 6**) and PD (**Figure 8**), Chao *et al.* (2008) suggest the use of a similarity profile $\{C_{qN}; q=0, 1, 2, \ldots, N\}$ to describe similarity across N assemblages. It is recommended that investigators calculate at least C_{0N}, C_{1N}, and C_{2N}; see Jost *et al.* (2011, p. 81) and **Figure 9** for examples.

Because the overlap measure C_{qN} is constructed from Hill numbers of order q, the magnitude of the undersampling bias depends on the order q. As with Hill numbers, there exists a nearly unbiased estimator only for $q=2$:

$$\hat{C}_{2N} = \frac{2\sum_{i=1}^{S}\sum_{j<k}(X_{ij}/n_j)(X_{ik}/n_k)}{(N-1)\sum_{i=1}^{S}\sum_{j=1}^{N}[X_{ij}(X_{ij}-1)/n_j(n_j-1)]} \qquad [37]$$

This measure quantifies the similarity of relative frequencies for abundant species. A bootstrap variance estimator and the associated confidence interval are provided in Chao *et al.* (2008). Satisfactory estimators for C_{0N} and C_{1N} that characterize the similarity of relative frequencies for rare species are still lacking.

Similarity Indices Based on Total Abundance of Shared Species

As explained earlier, the Horn overlap measure (eqn [30]) and the Morisita–Horn similarity measure (eqn [31]) match species relative abundances, *species-by-species*. Hence, the typical similarity indices assess a normalized probability that two randomly chosen individuals, one from each assemblage, belong to *the same* species. Another approach by Chao *et al.* (2005) is to consider a (normalized) probability that both individuals, one from each of the two assemblages, both belong to *any shared* species (and not necessarily to the same shared species).

Let U denote the total relative abundances associated with the shared species in Assemblage 1 and let V denote the total relative abundances of the shared species in Assemblage 2. The Jaccard abundance-based similarity index is $UV/(U+V-UV)$ and the Sørensen abundance-based similarity index is $2UV/(U+V)$. These two shared-abundance indices are called the Chao–Jaccard abundance and Chao–Sørensen abundance indices in the literature and in the software package *EstimateS* (Colwell, 2011) and SPADE (Chao and Shen, 2010). This is because, when all species are equally common, $U=S_{12}/S_1$ and $V=S_{12}/S_2$ and the Chao–Jaccard abundance and Chao–Sørensen abundance indices reduce, respectively, to the classic incidence-based Jaccard and Sørensen indices. These two measures yield a maximum value of 1 when all species are shared (i.e., no unique species in both assemblages; $U=V=1$). Also, all indices tend to a minimum value of 0 for completely distinct assemblages (i.e., no shared species in both assemblages).

One advantage of these measures is that the undersampling bias due to unseen, shared species can be evaluated and corrected. Chao *et al.* (2005) used the frequencies of observed rare, shared species to obtain an appropriate adjustment term for U and V to account for the effect of *unseen* shared species and thus remove most undersampling bias. Then the bias-corrected U and V estimators are substituted into the formulas to obtain Chao–Jaccard and Chao–Sørensen estimators. These measures are designed to be sensitive to rare shared species while still taking abundance into account, so they may increase sharply as more shared species are discovered. Because these measures match the total relative abundances of species shared between two assemblages, they are useful if the focus is to construct abundance-based *complementarity* (dissimilarity or distance) measures by subtracting each measure from one. This class of measures can also be extended to replicated incidence data; see Chao *et al.* (2005) for details.

Phylogenetic Similarity Indices

The classic Jaccard, Sørensen, and Morisita-Horn similarity measures all have their own phylogenetic generalizations. Most of the pioneering work was developed by microbial ecologists (Lozupone and Knight, 2005; Faith *et al.*, 2009). The phylogenetic Jaccard and Sørensen measures are based on Faith's total branch lengths and have formulas similar to their classic versions. The phylogenetic Sørensen index can be expressed as $2L_{12}/(L_1+L_2)$, 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 in the same time interval of interest (Lozupone and Knight, 2005). The phylogenetic Jaccard index takes the form of $L_{12}/(L_1+L_2-L_{12})$. When species relatedness is based on a simple Linnean taxonomic classification tree, L_1 and L_2 become the number of taxa in Trees 1 and 2, respectively, and L_{12} becomes the number of shared taxa in the pooled classification tree (Bacaro *et al.*, 2007). In these generalizations, "species" in the traditional indices are replaced by the total branch lengths (or the total number of nodes) in each assemblage. Also, "shared species" in the traditional indices are replaced by the total shared branch length (or the total number of shared nodes). In nearly all applications of these phylogenetic similarity indices, it is assumed that all species are observed; undersampling bias due to undetected species for these measures has not been discussed in this literature.

The classic Morisita–Horn measure has recently been generalized to its phylogenetic version (de Bello *et al.*, 2010). Let $\boldsymbol{B}$ denote the set consisting of all branches in the pooled assemblages in a specific time period of interest, and let the corresponding branch lengths in this set be $\{L_i;\, i\in\boldsymbol{B}\}$. Assume that, in the jth assemblage, a_{ij} denotes the total relative abundance descended from Branch i, $i\in\boldsymbol{B}$, $j=1, 2, \ldots, N$. The phylogenetic Morisita–Horn similarity for N assemblages is a generalization of eqn [35c] based on a normalized Rao's quadratic entropy:

$$S^{*}_{\mathrm{MH}}=\frac{2\sum_{i\in B}\sum_{j<k}L_i a_{ij}a_{ik}}{(N-1)[\sum_{i\in B}\sum_{j=1}^{N}L_i a_{ij}^2]} \qquad [38]$$

A nearly unbiased estimator for this measure is similar to that in eqn [37]. However, the measure in eqn [38] is valid only for ultrametric trees. Extension to nonultrametric trees requires further research.

Appendix

List of Courses

1. Community Ecology
2. Conservation Biology
3. Statistical Ecology

See also: Biodiversity, Definition of. Defining, Measuring, and Partitioning Species Diversity. Diversity, Molecular Level. Diversity, Taxonomic versus Functional. Functional Diversity Measures. Latitudinal Gradients of Biodiversity. Measurement and Analysis of Biodiversity. Nucleic Acid Biodiversity: Rewriting DNA and RNA in Diverse Organisms. Species Diversity, Overview

References

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.

Bacaro G, Ricotta C, and Mazzoleni S (2007) Measuring beta-diversity from taxonomic similarity. *Journal of Vegetation Science* 18: 793–798.

de Bello F, Lavergne S, Meynard CN, Lepš J, and Thuiller W (2010) The partitioning of diversity: Showing Theseus a way out of the labyrinth. *Journal of Vegetation Science* 21: 1–9.

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 (2005) Species estimation and applications. In: Kotz S, Balakrishnan N, Read CB, and Vidakovic B (eds.) *Encyclopedia of Statistical Sciences*, 2nd edn, pp. 7907–7916. New York: Wiley.

Chao A, Chazdon RL, Colwell RK, 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, 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, Colwell RK, Lin C-W, and Gotelli NJ (2009) Sufficient sampling for asymptotic minimum species richness estimators. *Ecology* 90: 1125–1133.

Chao A and Jost L (2011) Diversity measures. In: Hastings A and Gross L (eds.) *Encyclopedia of Theoretical Ecology*. Berkeley: University of California Press. In press.

Chao A, Jost L, Chiang SC, Jiang Y-H, and Chazdon RL (2008) A two-stage probabilistic approach to multiple-community similarity indices. *Biometrics* 64: 1178–1186.

Chao A and Shen T-J (2003) Nonparametric estimation of Shannon's index of diversity when there are unseen species. *Environmental and Ecological Statistics* 10: 429–443.

Chao A and Shen TJ (2004) Nonparametric prediction in species sampling. *Journal of Agricultural, Biological, and Environmental Statistics* 9: 253–269.

Chao A and Shen T-J (2010) SPADE: Species Prediction and Diversity Estimation. Program and User's Guide at http://chao.stat.nthu.edu.tw/softwareCE.html.

Coleman BD, Mares MA, Willig MR, and Hsieh Y-H (1982) Randomness, area, and species richness. *Ecology* 63: 1121–1133.

Colwell RK (2011) EstimateS: Statistical estimation of species richness and shared species from samples. Version 9. User's Guide and application published at: http://purl.oclc.org/estimates.

Colwell RK, Chao A, Gotelli NJ, *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 RK and Coddington JA (1994) Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B* 345: 101–118.

Colwell RK, Mao CX, and Chang J (2004) Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology* 85: 2717–2727.

Ellison AM, Gotelli NJ, Farnsworth EJ, and Alpert GD (2012) *A Field Guide to the Ants of New England*. New Haven, CT: Yale University Press.

Efron B and Tibshirani RJ (1993) *An Introduction to the Bootstrap*. London: Chapman and Hall.

Faith DP (1992) Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61: 1–10.

Faith DP, Lozupone CA, Nipperess D, and Knight R (2009) The cladistic basis for the phylogenetic diversity (PD) measure links evolutionary features to environmental gradients and supports broad applications of microbial ecology's "phylogenetic beta diversity" framework. *International Journal of Molecular Sciences* 10: 4723–4741.

Gotelli NJ and Colwell RK (2001) Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4: 379–391.

Gotelli NJ and Colwell RK (2011) Estimating species richness. In: Magurran A and McGill B (eds.) *Biological Diversity: Frontiers in Measurement and Assessment*, pp. 39–54. Oxford: Oxford University Press.

Hill MO (1973) Diversity and evenness: A unifying notation and its consequences. *Ecology* 54: 427–431.

Horn HS (1966) Measurement of "overlap" in comparative ecological studies. *American Naturalist* 100: 419–424.

Hurlbert SH (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 RL (2011) Compositional similarity and beta diversity. In: Magurran A and McGill B (eds.) *Biological Diversity: Frontiers in Measurement and Assessment*, pp. 66–84. Oxford: Oxford University Press.

Lennon JJ, Koleff P, Greenwood JJD, and Gaston KJ (2001) The geographical structure of British bird distributions: Diversity, spatial turnover and scale. *Journal of Animal Ecology* 70: 966–979.

Lozupone C and Knight R (2005) UniFrac: A new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* 71: 8228–8235.

MacArthur RH (1965) Patterns of species diversity. *Biological Reviews* 40: 510–533.

Magurran AE (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 (Biology)* 3: 65–80.

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.

Petchey OL and Gaston KJ (2002) Functional diversity (FD), Species richness and community composition. *Ecology Letters* 5: 402–411.

Pielou EC (1975) *Ecological Diversity*. New York: John Wiley.

Rao CR (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. *10th Annual Meeting of the Ecological Society of Japan* (Abstract), p.5.

Tokeshi M (1999) *Species Coexistence: Ecological and Evolutionary Perspectives*. Oxford: Blackwell.

Tuomisto H (this volume) Defining, measuring and partitioning species diversity. *Encyclopedia of Biodiversity*.

Vane-Wright RI, Humphries CJ, and Williams PM (1991) What to protect: Systematics and the agony of choice. *Biological Conservation* 55: 235–254.