Received: 7 September 2019 | Revised: 10 February 2020 | Accepted: 13 February 2020

DOI: 10.1111/1440-1703.12102

ECOLOGICAL RESEARCH WILEY

SPECIAL FEATURE

Approaches for general rules of biodiversity patterns in space and time

Quantifying sample completeness and comparing diversities among assemblages

Anne Chao[1] | Yasuhiro Kubota[2] | David Zelený[3] | Chun-Huo Chiu[4] | Ching-Feng Li[5] | Buntarou Kusumoto[2,6] | Moriaki Yasuhara[7] | Simon Thorn[8] | Chih-Lin Wei[9] | Mark J. Costello[10,11] | Robert K. Colwell[12,13]

[1]Institute of Statistics, National Tsing Hua University, Hsin Chu, Taiwan

[2]Faculty of Science, University of the Ryukyus, Nishihara, Japan

[3]Institute of Ecology and Evolutionary Biology, National Taiwan University, Taipei, Taiwan

[4]Department of Agronomy, National Taiwan University, Taipei, Taiwan

[5]School of Forestry and Resource Conservation, National Taiwan University, Taipei, Taiwan

[6]Royal Botanic Gardens, Kew, Richmond, UK

[7]School of Biological Sciences and Swire Institute of Marine Science, The University of Hong Kong, Pokfulam Road, Hong Kong

[8]Field Station Fabrikschleichach, Biocenter, University of Würzburg, Rauhenebrach, Germany

[9]Institute of Oceanography, National Taiwan University, Taipei, Taiwan

[10]School of Environment, University of Auckland, Auckland, New Zealand

[11]Nord University, Bodø, Norway

[12]Department of Ecology and Evolution, University of Connecticut, Storrs, USA

[13]University of Colorado Museum of Natural History, Boulder, USA

Correspondence
Anne Chao, Institute of Statistics, National Tsing Hua University, Hsin Chu 30043, Taiwan.
Email: chao@stat.nthu.edu.tw

Funding information
Taiwan Ministry of Science and Technology, Grant/Award Number: 107-2118-M007-006-MY3; Research Grants Council of Honk Kong, Grant/Award Numbers: HKU 709413P, HKU 17303115, HKU 173025518; University of Hong Kong, Grant/Award Number: 201111159140, 201611159053, 201711159057; JSPS KAKENHI, Grant/Award Number: 16H02755; University of the Ryukyus

Abstract

We develop a novel class of measures to quantify sample completeness of a biological survey. The class of measures is parameterized by an order $q \geq 0$ to control for sensitivity to species relative abundances. When $q = 0$, species abundances are disregarded and our measure reduces to the conventional measure of completeness, that is, the ratio of the observed species richness to the true richness (observed plus undetected). When $q = 1$, our measure reduces to the sample coverage (the proportion of the total number of individuals in the entire assemblage that belongs to detected species), a concept developed by Alan Turing in his cryptographic analysis. The sample completeness of a general order $q \geq 0$ extends Turing's sample coverage and quantifies the proportion of the assemblage's individuals belonging to detected species, with each individual being proportionally weighted by the $(q - 1)$th power of its abundance. We propose the use of a continuous profile depicting our proposed

This paper is dedicated to the memory of our coauthor Ching-Feng (Woody) Li, who passed away on November 29, 2019, after a courageous battle with lung cancer. We hope Woody is free, without pain, happily hiking somewhere around Chilai mountain in Taiwan.

measures with respect to $q \geq 0$ to characterize the sample completeness of a survey. An analytic estimator of the diversity profile and its sampling uncertainty based on a bootstrap method are derived and tested by simulations. To compare diversity across multiple assemblages, we propose an integrated approach based on the framework of Hill numbers to assess (a) the sample completeness profile, (b) asymptotic diversity estimates to infer true diversities of entire assemblages, (c) non-asymptotic standardization via rarefaction and extrapolation, and (d) an evenness profile. Our framework can be extended to incidence data. Empirical data sets from several research fields are used for illustration.

KEYWORDS

completeness, diversity, evenness, Hill numbers, sample coverage

1 | INTRODUCTION

The goal of many biological surveys is to quantify and compare biodiversity across multiple assemblages. In a typical individual-based biological survey, a sample of individuals is collected from an assemblage; each individual is identified to species, and the abundance (or frequency) of each species in the sample is recorded. However, due to practical limitations, it is virtually impossible to detect all species, especially in hyperdiverse assemblages with many rare species (Colwell & Coddington, 1994; Gotelli & Colwell, 2001, 2011; Magurran & McGill, 2011). In almost every biodiversity survey and monitoring project, some proportion of the species that are present in the assemblage fail to be detected, and thus the abundances of those undetected species remain unknown. Before comparing diversity across assemblages, we should first quantify sample completeness of a biological survey and assess the extent of undetected diversity.

Most ecologists intuitively think that sample completeness is measured by the ratio of the observed species in a sample to the true species richness (observed plus undetected) in the entire assemblage. Underlying this perspective is the conventional sense that all *species* are treated as equal while species abundances are completely disregarded. By contrast, if *individuals* are treated equally so that each species is weighted by its abundance, a widely used definition of sample completeness is the *sample coverage* (the proportion of the total number of individuals in an assemblage that belong to the species represented in the sample), a concept originally developed by Alan Turing in his cryptographic analysis during World War II. Turing developed the concept based on the observed frequencies in "samples" of intercepted Nazi code. While Turing never published his war-time statistical work, he gave permission to I. J. Good to publish it. The following presentation of the concept of sample coverage is based primarily on Good's papers (Good, 1953, 1983, 2000; Good & Toulmin, 1956). The concept and related topics have found many applications in various disciplines (as detailed by McGrayne, 2011). To the best of our knowledge, a unified approach that can integrate both a species-focused approach and an individual-focused approach to quantifying sample completeness is still lacking.

In this paper, we propose in Section 2 a novel class of sample completeness measures, parameterized by an order $q \geq 0$, where q is a number that determines the measures' sensitivity to species abundances. For a general order $q \geq 0$, our sample completeness measure extends Turing's concept of sample coverage to a generalized sample coverage in which each *species* is proportionally weighted by the qth power of its abundance, that is, each *individual* is proportionally weighted by the $(q - 1)$th power of its abundance. In the special case of $q = 0$, our sample completeness reduces to the conventional measure: the ratio of the observed species richness to the true species richness. When $q = 1$, it reduces to Turing's sample coverage (or simply "coverage"), with each species proportionally weighted by its abundance, or, equivalently, with each individual weighted by a constant. When $q = 2$, it represents a generalized sample coverage with each species being proportionally weighted by its squared species abundance (i.e., each individual being proportionally weighted by its species abundance); this measure thus is disproportionally sensitive to highly abundant species.

Our framework not only integrates the intuitive sense of sample completeness with Turing's concept of sample coverage, but also generalizes Turing's concept to a more general class of measures of order $q \geq 0$. Rather than using one or a few measures, we propose the use of a continuous profile that depicts our measures with respect to the order $q \geq 0$ to characterize the sample completeness of a biological survey. This approach makes it easy to visually compare completeness profiles of multiple assemblages. In most applications, a completeness profile for all values of q from $q = 0$ to $q = 2$ would be sufficient, as beyond $q = 2$ the profile generally stabilizes and changes only slowly.

In practice, the proposed sample-completeness profile needs to be estimated from sampling data. In Section 3, we develop an analytic estimator for the profile and test it through simulation experiments. A bootstrap method is proposed to obtain the associated confidence intervals, facilitating a comparison of completeness among multiple datasets. If all species are observed in a survey, then the estimated profile has a sample completeness of unity for all orders of $q \geq 0$, which occurs only when there are no singletons in the data. Otherwise, the estimated completeness profile generally increases with order q, a pattern that indicates an incomplete sample. We will use several examples to illustrate our estimated sample-completeness profiles.

Until fairly recently, how to quantify and compare biodiversity across assemblages was one of the most controversial issues in ecology (Magurran & McGill, 2011). Following a multi-author *Ecology* forum (Ellison, 2010), a consensus seems to have emerged that Hill numbers (Hill, 1973) should be the species diversity measure of choice. In his influential paper, Hill integrated species richness and abundance into a class of diversity measures that later came to be known as *Hill numbers*. Defined as the effective number of equally abundant species, this class of measures is parameterized by a diversity order q, the same order that we use for quantifying sample completeness in this paper. Thus, our framework links the concepts of sample completeness and diversity. Hill numbers for order $q \geq 0$ are all in units of "species" or "species equivalents" and include the three most widely used species diversity measures (species richness, Shannon diversity and Simpson diversity) as special cases of orders $q = 0$, 1 and 2, respectively. A diversity profile that depicts Hill numbers with respect to the order $q \geq 0$ conveys all the information in a species abundance distribution. Section 4.1 presents a brief review of Hill numbers.

Empirical or observed Hill number of any order (including species richness) based on sampling data is dependent on sample size and sample completeness (Chao et al., 2014) and thus cannot be used for comparing species diversity across multiple assemblages. To make fair diversity comparison among assemblages, some fundamental advances were made in the past decade, as outlined below and detailed in Sections 4.2–4.5.

1. An asymptotic analysis via statistical estimation of *true* diversities (Chao & Jost, 2015). This approach aims to compare asymptotic estimates of true diversities of entire assemblages. Here, an "asymptotic" value refers to the diversity estimate that would be reached when the sample size is hypothetically expanded to be large. However, sufficient data are required to accurately infer the true diversities. Whether data are sufficient can be determined by visually examining the estimated diversity accumulation curve with respect to sample size, that is, the sample-size-based rarefaction and extrapolation sampling curve of Hill numbers (Chao et al., 2014; Colwell et al., 2012); see Sections 4.2 and 4.3 for details.

2. A non-asymptotic standardization approach via coverage-based rarefaction and extrapolation. This approach aims to compare diversity estimates for equally complete samples, where sample completeness is measured by sample coverage (our completeness measure of order $q = 1$). When the data do not contain sufficient information to accurately infer the true diversity of an entire assemblage, we can infer the diversity for a standardized sample coverage, that is, a standardized fraction of the assemblage's individuals, and make fair diversity comparisons across multiple assemblages. Section 4.4 provides a brief review; see Colwell et al. (2012), Chao and Jost (2012) and Chao et al. (2014) for mathematical derivation and formulas.

3. Chao and Ricotta (2019) recently proposed linking evenness to diversity (Hill numbers) and developed five classes of evenness measures. One class of measures is based on the slopes of diversity profiles. That is, they considered the slope connecting two points with diversity orders 0 and any $q > 0$ in the Hill-number profile. The slope is then normalized to the range of [0, 1] and converted to the corresponding evenness measure. Their measures can be applied to compare evenness when species richness is not fixed across assemblages; see Section 4.5.

Within the common framework of Hill numbers, we propose in Section 5 a unified approach that integrates the development on sample completeness (new in this paper) with all the advances mentioned above. Our proposed approach comprises a four-step procedure to assess for each assemblage (a) the sample completeness profile proposed in this paper, (b) asymptotic diversity estimates to infer the true diversities of entire assemblages, (c) non-asymptotic coverage-based rarefaction and extrapolation and (d) an evenness profile derived from the slopes of the diversity profile. See Colwell and Chao (2020) for a brief guide to the history and state-of-the-art in diversity statistics with an application of the procedure to an archeological dataset. We applied our methodologies here to four examples from contrasting fields so as to demonstrate the wide applicability of our approach.

Gotelli and Colwell (2001) distinguished two types of biological survey data: abundance data (in which individuals are randomly selected) and incidence data (in which sampling units are randomly selected). For the latter, the sampling unit is often a trap, net, quadrat, plot, or timed survey and only species' occurrence (detection and non-detection) records in each sampling unit are required. Colwell et al. (2012), Chao et al. (2014) and Chao and Colwell (2017) demonstrated that replicated incidence data support statistical approaches to the biological inference that are just as powerful as the corresponding abundance-data-based approaches. Most importantly, analyses based on multiple incidence data are less sensitive to clustering or aggregation of individuals, compared to those based on abundance data (Colwell et al., 2012; Colwell, Mao, & Chang, 2004). For example, species abundances of woody plants are often recorded in each of the selected plots or quadrats in vegetation surveys. Due to spatial aggregation, individual plants cannot be modeled as *independent* sampling units and thus the basic assumption for the abundance-data model is not fulfilled. In this case, it is preferable to first convert species abundance records in each plot to incidence data; each plot can then be treated as a sampling unit to satisfy the basic independence assumption of sampling units in the incidence-data model. Here, we also extend our framework of sample-completeness and diversity comparison to deal with replicated incidence data.

In Section 6, two abundance-based datasets (fossil ostracods and spiders), and two incidence-based datasets (woody plants and stony corals) are analyzed to illustrate our suggested procedures; interpretations are also drawn from the results of each analysis. Readers who are interested only in applications may move directly to Section 5 (*An integrated four-step procedure*) and Section 6 (*Empirical examples*) for real data analysis. Some relevant issues are discussed in Section 7 and a conclusion is given the final section. Mathematical details and simulation results are provided in the Supporting Information.

2 | SAMPLE COMPLETENESS PROFILES

2.1 | Theoretical framework for abundance data

Assume that there are S species in the focal assemblage with species relative abundances $(p_1, p_2, \ldots, p_S)$, $\sum_{i=1}^{S} p_i = 1$. Suppose a *reference sample* of n individuals is selected, with replacement, from the assemblage. Let X_i denote the observed abundance/frequency of the ith species in the sample, $i = 1, 2, \ldots, S$, $\sum_{X_i \ge 1} X_i = n$. A commonly-used model specifies that the sample frequencies $(X_1, X_2, \ldots, X_S)$ follow a multinomial distribution with cell total n and cell probabilities $(p_1, p_2, \ldots, p_S)$. The marginal distribution for sample frequency X_i follows a binomial distribution, characterized by n and probability p_i. Only those species with frequency $X \ge 1$ are detected in the sample; those species with abundance $X = 0$ in the sample remain undetected and are therefore not included in the data.

Based on species frequencies $(X_1, X_2, \ldots, X_S)$, the *abundance-based frequency count* f_r, $r = 0, 1, \ldots, n$, is defined as the number of species each represented by exactly r individuals in the sample. Thus, f_0 is the number of undetected species, f_1 is the number of "singletons" (those species that are represented by exactly one individual in the sample), and f_2 is the number of "doubletons" (those that are represented by exactly two individuals in the sample). In his cryptanalysis, Turing determined that singletons and doubletons contain most of the information about undetected code elements. In our approach, singletons and doubletons also play an important role in our inference of sample completeness.

Given the true species relative abundances $(p_1, p_2, \ldots, p_S)$ of the S species in the assemblage, let the qth power sum be denoted as ${}^{q}\lambda = \sum_{i=1}^{S} p_i^q$ with ${}^{0}\lambda = S$ and ${}^{1}\lambda = 1$. The theoretical sample completeness of order q based on sample frequencies $(X_1, X_2, \ldots, X_S)$ is defined as the proportion of the detected qth power sum with respect to the true qth power sum of the entire assemblage, including undetected species. That is, we define the theoretical sample completeness of the qth order as

$$ {}^{q}C = \frac{{}^{q}\lambda_{\text{detected}}}{{}^{q}\lambda} = \frac{\sum\limits_{i \in \text{detected}} p_i^q}{\sum_{i=1}^{S} p_i^q} = \frac{\sum_{i=1}^{S} p_i^q \, I(X_i > 0)}{\sum_{i=1}^{S} p_i^q}, \quad q \ge 0, \tag{1} $$

where $I(\cdot)$ is an indicator function that equals 1 when the specified condition is true and 0 otherwise. To gain an intuitive meaning of our measures of sample completeness, consider the following three special cases. See the upper half of Table 1 (second column) for the theoretical formula of a general order $q \ge 0$ and the three special cases ($q = 0$, 1 and 2).

(1) When $q = 0$, the theoretical measure reduces to the proportion of species that have been observed, that is, ${}^{0}C = S_{\text{obs}}/S$, where S_{obs} denotes the number of observed species in the sample. This measure expresses the conventional sense of sample completeness familiar to most ecologists; here, species abundances are completely disregarded. This zero-order measure quantifies the sample completeness when all species are treated equally and have a constant weight, or, equivalently, when each

TABLE 1 Theoretical formulas and analytic estimators for sample completeness measures of order $q \geq 0$ and three special cases ($q = 0$, 1 and 2) for (a) abundance data and (b) sampling-unit-based replicated incidence data. See the text and footnotes for notation and details

Order q	Theoretical formula	Analytic estimator
(a) Abundance data		
$q \geq 0$	${}^{q}C = \frac{{}^{q}\lambda_{\text{detected}}}{{}^{q}\lambda} = \frac{\sum_{i \in \text{detected}} p_i^q}{\sum_{i=1}^{S} p_i^q}$	${}^{q}\hat{C} = \frac{{}^{q}\hat{\lambda}_{\text{detected}}}{{}^{q}\hat{\lambda}} = 1 - \frac{f_1}{n}\left[\frac{A^{q-1}(1-A)}{{}^{q}\hat{\lambda}}\right]$
$q = 0$	${}^{0}C = \frac{S_{\text{obs}}}{S}$	${}^{0}\hat{C} = \frac{S_{\text{obs}}}{\hat{S}_{\text{Chao1}}}$
$q = 1$	${}^{1}C = \sum_{i \in \text{detected}} p_i$	${}^{1}\hat{C} = 1 - \frac{f_1}{n}(1-A)$
$q = 2$	${}^{2}C = \frac{\sum_{i \in \text{detected}} p_i^2}{\sum_{i=1}^{S} p_i^2}$	${}^{2}\hat{C} = 1 - \frac{f_1}{n}\left[\frac{A(1-A)}{\sum_{X_i \geq 2} X_i(X_i-1)/[n(n-1)]}\right]$
(b) Replicated incidence data		
$q \geq 0$	${}^{q}C = \frac{{}^{q}\Phi_{\text{detected}}}{{}^{q}\Phi} = \frac{\sum_{i \in \text{detected}} \pi_i^q}{\sum_{i=1}^{S} \pi_i^q}$	${}^{q}\hat{C} = \frac{{}^{q}\hat{\Phi}_{\text{detected}}}{{}^{q}\hat{\Phi}} = 1 - \frac{Q_1}{T}\left[\frac{B^{q-1}(1-B)}{{}^{q}\hat{\Phi}}\right]$
$q = 0$	${}^{0}C = \frac{S_{\text{obs}}}{S}$	${}^{0}\hat{C} = \frac{S_{\text{obs}}}{\hat{S}_{\text{Chao2}}}$
$q = 1$	${}^{1}C = \frac{\sum_{i \in \text{detected}} \pi_i}{\sum_{i=1}^{S} \pi_i}$	${}^{1}\hat{C} = 1 - \frac{Q_1}{T}\frac{(1-B)}{\sum_{Y_i \geq 1} Y_i/T} = 1 - \frac{Q_1}{U}(1-B)$
$q = 2$	${}^{2}C = \frac{\sum_{i \in \text{detected}} \pi_i^2}{\sum_{i=1}^{S} \pi_i^2}$	${}^{2}\hat{C} = 1 - \frac{Q_1}{T}\left[\frac{B(1-B)}{\sum_{Y_i \geq 2} Y_i(Y_i-1)/[T(T-1)]}\right]$

Note: (1) See the Supporting Information for the formulas of ${}^{q}\hat{\lambda}$ (Equation S1.6), ${}^{q}\hat{\lambda}_{\text{detected}}$ (Equation S1.9), ${}^{q}\hat{\Phi}$ (Equation S2.6), and ${}^{q}\hat{\Phi}_{\text{detected}}$ (Equation S2.9), where ${}^{q}\lambda = \sum_{i=1}^{S} p_i^q$ and ${}^{q}\Phi = \sum_{i=1}^{S} \pi_i^q$ denote, respectively, the qth power sum for abundance and incidence data.

(2) A (for abundance data): the estimated mean relative frequency of singletons; B (for incidence data): the estimated mean detection probability in any sampling unit of unique species.

$$A = \begin{cases} 2f_2/[(n-1)f_1 + 2f_2], & \text{if } f_2 > 0; \\ 2/[(n-1)(f_1-1) + 2], & \text{if } f_2 = 0,\ f_1 \neq 0; \\ 1, & \text{if } f_2 = f_1 = 0. \end{cases} \quad B = \begin{cases} 2Q_2/[(T-1)Q_1 + 2Q_2], & \text{if } Q_2 > 0; \\ 2/[(T-1)(Q_1-1) + 2], & \text{if } Q_2 = 0, Q_1 > 0; \\ 1, & \text{if } Q_1 = Q_2 = 0. \end{cases}$$

(3) $U = \sum_{Y_i \geq 1} Y_i$ denotes the total number of incidences based on detection/non-detection records of T sampling units.

individual is weighted by $1/p$, the inverse of its species relative abundance. Therefore, this measure is disproportionally sensitive to rare species, compared to measures with order $q > 0$.

(2) When $q = 1$, the measure ${}^{1}C$ reduces to the sum of the relative abundances of the detected species, or, equivalently, the fraction of the assemblage's individuals that belong to the detected species. This is the concept of Turing's sample coverage (Good, 1953, 2000), which quantifies sample completeness when all individuals are treated equally. The weight for every individual is the same, regardless of species, so that a species' weight is proportional to its abundance, without disproportionally favoring either abundant or rare species.

(3) When $q = 2$, the measure quantifies the fraction of the total number of individuals in the assemblage that belong to the detected species, with each species being proportionally weighted by its squared species relative abundance, or, equivalently, with each individual being proportionally weighted by its species abundance. Thus, the measure ${}^{2}C$ is disproportionally sensitive to highly abundant species. In most surveys, highly abundant species would be detected in any sample; thus, the second- and higher-order measures typically yield values very close to unity.

The sample completeness measure of any order $q \geq 0$ quantifies a generalized sample coverage, that is, the proportion of the total number of individuals in the assemblage belonging to detected species, with each species being proportionally weighted by p^q, the qth power of its species abundance. Equivalently, each individual is proportionally weighted by p^{q-1}. Our measures of orders $q > 1$ are disproportionately sensitive to highly abundant species, whereas the measures of orders $q < 1$ are disproportionately sensitive to rare species. The measure of order $q = 1$ reduces to Turing's sample coverage, as described above.

Although $(p_1, p_2, ..., p_S)$ is modeled in our theory as species relative abundances, our derivation is also valid under a more general model in which $(p_1, p_2, ..., p_S)$ represent species detection probabilities. Generally, the detection probability for any individual is a combination of species abundance and the individual's detectability, which is determined by many factors such as the individual's color, size, (and for animals) vocalizations and movement patterns. If individuals of all species are assumed to have identical detectability, then species detection probability reduces to species relative abundance.

2.2 | Theoretical framework for incidence data

Replicated incidence data for a *reference sample* consist of incidence or occurrence (detection/non-detection) records for a set of T sampling units. The detection or non-detection of each species within each sampling unit is recorded to form a species-by-sampling-unit incidence matrix with S rows and T columns. The (i, j) element of the incidence data matrix is 1 if species i is detected in the jth sampling unit, and 0 if it is undetected in that sampling unit. The model assumes that the ith species is detected in any sampling unit with its own unique *incidence* (or *detection*) *probability* π_i. For example, in quadrat sampling, the incidence probability of a species represents the proportion of the number of quadrats in which that species can be detected. Each incidence is analogous to an "individual" and incidence probability π_i is analogous to p_i in the abundance-data model. Here, $\sum_{i=1}^{S}\pi_i$ may be greater than unity because it represents the expected number of species that are detected in any sampling unit. The row sum of an incidence matrix, Y_i, denotes the *incidence frequency* of species i, where $i = 1, 2,..., S$. A species has been detected in incidence data if that species is detected in at least one sampling unit (i.e., $Y > 0$). Species present in the assemblage but not detected in any sampling unit yield $Y = 0$.

Given a set of detection probabilities $(\pi_1, \pi_2, ..., \pi_S)$, the marginal distribution for the incidence-based frequency Y_i for the ith species follows a binomial distribution characterized by T and the detection probability π_i. The model is analogous to the binomial model for abundance data, that is, the sample frequency X_i for the ith species follows a binomial distribution characterized by n and the species relative abundance p_i. This analogy explains why statistical inferences for abundance data and incidence data are, in principal, parallel, although results may differ substantially depending upon spatial patterning of data.

Let Q_k denotes the *incidence-based frequency counts*, that is, the number of species that are detected in exactly k sampling units, $k = 0, 1,..., T$. In other words, the count Q_k is the number of species each represented exactly k times in the incidence matrix. Here, Q_k is analogous to f_k in the abundance data: Q_0 represents the number of species present in the assemblage but not detected in any of the T sampling units, Q_1 represents the number of *unique* species (those that are each detected in only one sampling unit) and Q_2 represents the number of *duplicate* species (those that are each detected in exactly two sampling units). Denote U as the total number of incidences recorded in the T sampling units; U is analogous to the sample size n in abundance data. However, n is fixed for abundance data whereas U varies according to the data and can be expressed as $U = \sum_{k=1}^{T} kQ_k = \sum_{i=1}^{S} Y_i$.

Denote the qth power sum of species detection probabilities $(\pi_1, \pi_2, ..., \pi_S)$ as ${}^q\Phi = \sum_{i=1}^{S}\pi_i^q$ with ${}^0\Phi = S$ and ${}^1\Phi = \sum_{i=1}^{S}\pi_i$. As with abundance data, the theoretical sample completeness of order q based on incidence-based sample frequencies $(Y_1, Y_2, ..., Y_S)$ is defined as the proportion of the detected qth power sum. That is, the theoretical sample completeness of the qth order for incidence data is defined as

$$ {}^qC = \frac{{}^q\Phi_{\text{detected}}}{{}^q\Phi} = \frac{\sum\limits_{i \in \text{detected}} \pi_i^q}{\sum_{i=1}^{S}\pi_i^q} = \frac{\sum_{i=1}^{S}\pi_i^q\, I(Y_i > 0)}{\sum_{i=1}^{S}\pi_i^q}, \quad q \geq 0. \tag{2} $$

See the lower half of Table 1 (the second column) for the theoretical formulas of a general order $q \geq 0$ and three special cases ($q = 0$, 1 and 2). Consider the following three special cases while noting that all interpretations are parallel to those for abundance data:

(1) When $q = 0$, the measure reduces to ${}^0C = S_{obs}/S$, which conforms to the conventional sense of sample completeness; here species incidence-based frequencies are disregarded. As with abundance data, the zero-order measure quantifies the sample completeness when all species are treated equally. In other words, each incidence is weighted by $1/\pi$, and thus is disproportionally sensitive to infrequent species, compared to measures with order $q > 0$.

(2) When $q = 1$, the measure ${}^1C$ quantifies the proportion of the total number of incidences/occurrences belonging to detected species (i.e., species detected in at least one of the T sampling units). This measure generalizes the Good-Turing concept to incidence data and quantifies sample completeness when all incidences/occurrences are treated equally. That is, the weight for any incidence is the same, regardless of species, so that a species' weight is proportional to its detection probability, without disproportionally favoring either frequent or infrequent species.

(3) When $q = 2$, the measure ${}^2C$ quantifies the fraction of the total number of incidences that belong to detected species, with each species being proportionally weighted by its squared detection probability. The

measure ${}^2C$ disproportionally favors highly frequent species. Since such species would be detected in at least one sampling unit, the second- and higher-order measures typically yield values close to unity.

The sample completeness measure of any order $q \geq 0$ quantifies a generalized incidence-based sample coverage, that is, the proportion of the total number of incidences belonging to detected species, with each incidence being proportionally weighted by π^{q-1}. Measures of orders $q > 1$ are disproportionately sensitive to the highly frequent species, whereas measures of orders $q < 1$ are disproportionately sensitive to the infrequent species. The measure of order $q = 1$ reduces to Turing's incidence-based sample coverage, which weights all species by their detection probabilities, without favoring either frequent or infrequent species.

3 | ESTIMATING SAMPLE-COMPLETENESS PROFILES

3.1 | Abundance data

In practice, the true species richness S and the relative abundances $(p_1, p_2, ..., p_S)$ of those S species in Equation (1) for abundance data are unknown. To assess sample completeness defined in Equation (1) from sampling data, we need to estimate not only the qth power sum ${}^q\lambda = \sum_{i=1}^{S} p_i^q$ (the denominator in Equation 1), but also the detected qth power sum ${}^q\lambda_{\text{detected}}$ (the numerator in Equation 1). Chao and Jost (2015) provided an analytic estimator ${}^q\hat{\lambda}$ of ${}^q\lambda$ via estimation of the slopes of the corresponding species accumulation curve. In Data S1, we provide a brief review of their estimation procedures and the formula for their estimator ${}^q\hat{\lambda}$ (Equation S1.6); we also present the derivation of the proposed ${}^q\hat{\lambda}_{\text{detected}}$ (Equation S1.9), leading to our estimator ${}^q\hat{C}$ for any order $q \geq 0$. See the upper half of Table 1 (the third column) for the estimator of a general order q and for the three special cases ($q = 0$, 1 and 2).

In the special case of $q = 0$, the estimator of ${}^0C = S_{\text{obs}}/S$ turns out to be ${}^0\hat{C} = S_{\text{obs}}/\hat{S}_{\text{Chao1}}$, which is equivalent to replacing true species richness in the formula by the Chao1 species richness estimator (Chao, 1984). As shown by Chao et al. (2017), a simple sufficient condition for the Chao1 richness estimator to be nearly unbiased is that *rare* species (specifically, singletons and undetected species) have approximately homogeneous abundances; in this case, the abundant species could be highly heterogeneous without affecting the estimation result. When rare species are heterogeneous in their abundances, the Chao1 estimator is theoretically a lower bound of true species richness. In this case, the estimated proportion of detected species exhibits a positive bias when sample size is not sufficiently large to detect all species; see the *Simulation* section in the Supporting Information for numerical results.

Contrary to most people's intuition, Turing (Good, 1953) showed that sample coverage (our measure of order $q = 1$) can be accurately and efficiently estimated using information contained in the sample itself. Turing's famous estimator of sample coverage is $1 - f_1/n$ (the complement of the proportion of singletons). Our sample coverage estimator (see Table 1), originally proposed in Chao and Jost (2012), represents a slightly-modified and more accurate version of Turing's estimator. Because of its good statistical properties, the concept of sample coverage has been used to objectively quantify sample completeness in many biodiversity studies and has been standardized to compare diversity among assemblages. (e.g., Chao et al., 2014; Chao & Jost, 2012).

3.2 | Incidence data

As with abundance data, we need to estimate the qth power sum ${}^q\Phi = \sum_{i=1}^{S} \pi_i^q$ (the denominator in Equation 2) and the detected qth order sum ${}^q\Phi_{\text{detected}}$ (the numerator in Equation 2) for all orders $q \geq 0$. Chao and Jost (2015, their Appendix S7) derived an estimator ${}^q\hat{\Phi}$ via estimation of the slopes of the incidence-based species accumulation curve. In Data S2, we briefly review their estimation procedures, present the formula for their estimator ${}^q\hat{\Phi}$ (Equation S2.6) and derive the proposed estimator ${}^q\hat{\Phi}_{\text{detected}}$ (Equation S2.9). Our estimator ${}^q\hat{C}$ of sample completeness of order q is the ratio of ${}^q\hat{\Phi}_{\text{detected}}$ and ${}^q\hat{\Phi}$. The resulting formula for any order $q \geq 0$ and the three special cases ($q = 0$, 1 and 2) are presented in the lower half of Table 1 (the third column).

In the special case of $q = 0$, our estimator of ${}^0C = S_{\text{obs}}/S$ turns out to be ${}^0\hat{C} = S_{\text{obs}}/\hat{S}_{\text{Chao2}}$, which is equivalent to replacing true species richness in the formula by the Chao2 species richness estimator (Chao, 1987). Chao and Colwell (2017) concluded that even if frequent species are highly heterogeneous in detection probability, the Chao2 estimator is a nearly unbiased estimator if *infrequent* species (specifically, uniques and undetected species) have approximately homogeneous detection probabilities. Otherwise, it is theoretically a lower bound, and thus the estimated proportion of detected species richness becomes an upper bound. When $q = 1$, our sample coverage estimator represents an extension of Turing's estimator to incidence data. Chao and Jost (2012) and Chao et al. (2014) adopted this sample coverage as a standardization criterion to compare diversity among assemblages for incidence data.

3.3 | Estimated sample-completeness profile

For each type of data (abundance or incidence), we propose the use of the estimated sample completeness profile, which depicts the estimator ${}^q\hat{C}$ as a function of q, $q \geq 0$. In practice, a completeness profile is plotted for all values of q from $q = 0$ to $q = 2$, beyond which the profile generally stabilizes. In Data S1, a bootstrap method to obtain the associated confidence intervals is sketched (fully developed in Chao et al., 2014, their Appendix G). Generally, for any fixed order q, if the 95% confidence intervals do not overlap, then significant difference at a level of 5% is guaranteed. However, overlapped intervals do not guarantee non-significance (Colwell et al., 2012) and rigorous statistical tests should be performed to determine whether a difference is statistically significant.

To instigate the performance of the proposed estimators of sample completeness, we report some simulation results based on abundance data (Data S3). The corresponding procedures and conclusions for incidence data are generally similar. In our simulation, species abundance data were simulated from five abundance models and two sample sizes. Generally, our estimated profiles are nearly unbiased for order $q \geq 1$. For $q < 1$, although positive bias exists when rare species are heterogeneous in their abundances, our estimated profiles provide very informative upper bounds for sample completeness; see Figure S3.1.

4 | HILL NUMBERS AND SOME RECENT ADVANCES

4.1 | Diversity (Hill numbers) profile

As indicated in the Introduction, we adopt Hill numbers to quantify the species diversity of an assemblage. For abundance data, the diversity of order q, ${}^qD$, can be expressed as a function of the qth power sum, that is, ${}^qD = \left({}^q\lambda\right)^{1/(1-q)} = \left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}$, $q \geq 0, \quad q \neq 1$, where ${}^q\lambda = \sum_{i=1}^{S} p_i^q$ denotes the qth power sum of species relative abundances, as previously defined in Equation (1). This class of diversities quantifies the effective number of equally abundant species in an assemblage. In the special case of $q = 0$, the Hill number reduces to species richness, which counts all *species* equally without regard to their relative abundances. For $q = 1$, the diversity of order $q = 1$, referred to as Shannon diversity (Chao et al., 2014), is the limit of ${}^qD$ as the order q tends to 1, that is, the exponential of Shannon entropy. The measure for $q = 1$ counts all *individuals* equally and thus weighs species in proportion to their abundances; the measure ${}^1D$ can be interpreted as the effective number of abundant species in the assemblage. The measure for $q = 2$, referred to as Simpson diversity, is expressed as the inverse of the Simpson concentration index; the measure ${}^2D$ discounts all but the highly abundant species and can be interpreted as the effective number of highly abundant species in the assemblage; see Hill (1973) and Chao et al. (2014) for a recent review.

Chao et al. (2014) defined the incidence-based species diversity of order q as the Hill numbers based on species *relative* detection probabilities for any occurrence record. As previously defined in *Theoretical framework for incidence data* (Section 2.2), let $0 < \pi_i < 1$ denote the detection probability of the ith species in any sampling unit. The *relative* detection probabilities become $\psi_i = \pi_i / \sum_{j=1}^{S} \pi_j = \pi_i / {}^1\Phi$, $i = 1, 2, \ldots, S$, where ${}^q\Phi = \sum_{j=1}^{S} \pi_j^q$ denotes the qth power sum of species detection probabilities, as defined in Equation (2). Here, ψ_i can also be interpreted as the probability that any detected incidence is classified to the ith species. By analogy to the case of abundance data, Hill numbers based on the probability set $(\psi_1, \psi_2, \ldots, \psi_S)$ are formulated as ${}^qD = \left(\sum_{i=1}^{S} \psi_i^q\right)^{1/(1-q)} = \left[{}^q\Phi / \left({}^1\Phi\right)^q\right]^{1/(1-q)}$, $q \neq 1$, and ${}^1D = \exp\left(-\sum_{i=1}^{S} \left(\pi_i / {}^1\Phi\right) \log\left(\pi_i / {}^1\Phi\right)\right)$. This class of measures quantifies the effective number of equally frequent species. For $q = 0$, this measure reduces to species richness, and the measures of $q = 1$ and $q = 2$ can be interpreted respectively as the effective number of frequent and highly frequent species in the assemblage.

4.2 | Rarefaction and extrapolation by sample size

It is well known that observed species richness based on sampling data is highly dependent on sample size (Colwell & Coddington, 1994). Here, sample size means the number of individuals for abundance data, or the number of sampling units for incidence data. Biologists have long recognized that species counts in samples cannot be used for comparing species richness across multiple assemblages. The traditional approach is to apply rarefaction to down-sample the larger samples until they contain the same sample size, but this necessarily results in throwing away some data in larger samples. To avoid discarding data, Colwell et al. (2012) developed for species richness a size-based rarefaction and extrapolation curve that depicts the estimated richness as a function of sample size. The curve can be rarefied to smaller sample sizes or extrapolated to larger sample sizes. Chao et al. (2014)

extended the rarefaction and extrapolation method to Hill numbers specifically for the three orders $q = 0$, 1 and 2. For species richness, the sample size can be extrapolated at most to double the reference sample size. For Hill numbers of orders $q = 1$ and 2, if data are not too sparse, the extrapolation can be reliably extrapolated to infinity.

4.3 | Asymptotic diversity estimates

An asymptotic approach refers to the comparison of the asymptotic estimates of true diversity profiles of entire assemblages. Our asymptotic diversity profile was developed by Chao and Jost (2015) via the analytic estimator of the qth power sum described earlier. The resulting formulas are given in the Supporting Information; see Equation (S1.7) for abundance data, and Equation (S2.7) for incidence data. They also developed a bootstrap method to obtain the associated confidence intervals.

As indicated in the Introduction, sufficient data are required to infer the true diversity, otherwise, the asymptotic estimates obtained from incomplete data may be subject to some bias. The above size-based sampling curve can be used to visually determine whether our asymptotic estimates can reliably infer true diversities. When the sample is extrapolated to double the size of the observed sample, if the rarefaction and extrapolation curve stabilizes and levels off (equivalently, the terminal slope tends to vanish), then Chao and Jost (2015) asymptotic estimates can be used to infer entire assemblages. This typically happens for Hill numbers of order $q \geq 1$. By contrast, if the curve is still increasing (typically for species richness), then the asymptotic estimator represents only a lower bound and thus exhibits negative bias, because there can always be some vanishingly rare species yet to be revealed. It turns out that species richness ($q = 0$) is the most difficult parameter to estimate. In the approach of Chao and Jost (2015), the asymptotic estimates of species richness for abundance and incidence data turn out to be, respectively, the Chao1 and Chao2 richness estimators (Chao, 1984, 1987). As stated earlier, these estimators generally are lower bounds.

4.4 | Rarefaction and extrapolation by sample coverage

When the data do not contain sufficient information to accurately infer the true diversity of an entire assemblage, Chao and Jost (2012) and Chao et al. (2014) advocated the use of a non-asymptotic standardization approach via coverage-based rarefaction and extrapolation with Hill numbers. The coverage-based sampling curve depicts diversity estimates as a function of sample coverage. This approach aims to compare diversity estimates for equally complete samples. They indicated that rarefaction and extrapolation to a given degree of sample coverage were better able to judge the magnitude of the differences in richness among assemblages, and ranked assemblages more efficiently, compared to traditional rarefaction and extrapolation to equal sample sizes. Because the terminal slope of a size-based rarefaction curve equals the coverage deficit (i.e., one minus coverage), standardizing coverage is equivalent to equalizing the slope of size-based sampling curves. Consequently, to standardize to a fixed value of coverage, more diverse assemblages require more sample sizes or sampling efforts; see Chao and Jost (2012, their figure 2) for an example and see Chao and Jost (2015, p. 52, their table 1) for methodologies and formulas.

For severely under-sampled assemblages, although the true diversities cannot be accurately assessed due to insufficient data, we can at least infer diversity for a standardized coverage or fraction of the assemblage's individuals and make fair diversity comparisons based on the resulting diversity. The inference can be made up to a maximum coverage or maximum fraction, denoted as $C_{\max}$. The value of $C_{\max}$ is selected as the minimum among the coverage values for samples extrapolated to double the size of the reference sample. That is, each sample is first extrapolated to double the reference sample size. Then the maximum assemblage fraction that we can accurately infer, $C_{\max}$, is the minimum among the coverage values obtained from those extrapolated samples. For any standardized coverage up to $C_{\max}$, diversity and evenness estimates (as discussed in the next subsection) for the standardized assemblage fraction can then be assessed and compared across assemblages.

4.5 | An evenness profile

Compared to diversity, quantifying evenness (or unevenness) among species abundances is an even more extensively discussed issue; see Chao and Ricotta (2019) for a recent review. When species abundances are completely even, the diversity profile is a horizontal line at the level of species richness. Otherwise, the profile is theoretically a decreasing function of order q. Therefore, the steepness of its slope reflects the unevenness of species abundances. When species richness is fixed, the more uneven the distribution of species abundances, the more steeply the profile declines; see figure 6 of Gotelli and Chao (2013) for an example. Chao and Ricotta (2019) considered the slope connecting two points with diversity orders 0 and any $q > 0$ in the Hill-number profile. The slope is then normalized to the range of [0, 1] to adjust for the effect of differing species richness. The resulting class of evenness measure of order q is expressed as $^qE = (^qD - 1)/(S - 1)$, for $q > 0$. (For $q = 0$, abundances are disregarded, so it is not meaningful to evaluate evenness.) They proposed quantifying evenness through a continuous profile which depicts evenness ^qE as a function of diversity order $q > 0$. This evenness profile can be applied to compare evenness even if species richness is not fixed across assemblages. Parallel arguments and the same evenness profile can be obtained for incidence data.

Hill (1973) proposed a definition of evenness as the ratio of diversity to species richness, that is, $^qD/S$. Although this ratio is only slightly different from the measure ^qE given above, Jost (2010) pointed out that this ratio ranges from $1/S$ to 1 because diversity takes values between 1 and S. Since the range is a function of richness, the ratio cannot be used for comparing the evenness of two assemblages with a different number of species. To remove the range's dependence on richness, Jost (2010) normalized the ratio to establish a class of evenness measures in the range of [0, 1]. The resulting class of evenness measures of order q is identical to the measure ^qE.

A widely used evenness measure is Pielou's J' (Pielou, 1966) which is expressed as $J' = \log(^1D)/(\log S) = H/(\log S)$, where H denotes Shannon entropy. Note that both J' and ^qE are functions of species richness S and diversity ^qD ($q > 0$). As discussed in the preceding subsection, in most applications, we can obtain only a lower bound of the true species richness for each assemblage. Thus, the "true" evenness of the entire assemblage cannot be accurately estimated from incomplete sampling data. Like our non-asymptotic analysis for species richness, evenness can be evaluated and compared among assemblages only if based on a standardized assemblage fraction (Chao & Ricotta, 2019; Jost, 2010). For all our examples in Section 5, we present evenness values for the coverage value of $C_{\max}$ (i.e., minimum coverage of samples extrapolated to double the size of the reference sample) defined earlier in Section 4.4.

5 | AN INTEGRATED FOUR-STEP PROCEDURE

We suggest the following four steps as guidelines to assess sample completeness and compare diversity across assemblages. This four-step procedure links sample completeness, diversity estimation, rarefaction and extrapolation, and evenness in a fully integrated approach.

Step 1. Assessment of sample completeness profile

If the estimated sample completeness profile is a horizontal line at the level of unity for all orders of $q \geq 0$, then the survey is complete, implying there is no undetected diversity. In most applications, the estimated profile increases with order q, revealing the existence of undetected diversity. The sample completeness value for $q = 0$ provides an upper bound for the proportion of observed *species*; its complement represents a lower bound for the proportion of undetected species. If data are not sparse, then the sample completeness value for $q = 1$ accurately measures the proportion of an assemblage's *individuals* belonging to detected species; and the values for $q \geq 2$ typically are very close to unity, signifying that almost all highly abundant species (for abundance data) or highly frequent species (for incidence data) had been detected in the reference sample.

Step 2. Size-based rarefaction and extrapolation analysis and the asymptotic diversity profile for $0 \quad q \quad 2$

(*Step 2a*). First examine the pattern of each size-based rarefaction and extrapolation sampling curve up to double the reference sample size for $q = 0$, 1 and 2. If the curve stays at a fixed level (this often occurs for the measures of $q = 1$ and 2), then our asymptotic estimate can be used to accurately infer the true diversity of the entire assemblage. Otherwise, our asymptotic diversity estimate represents only a lower bound.

(*Step 2b*). When the true diversity can be accurately inferred, the extent of undetected diversity within each dataset is obtained by comparing the estimated asymptotic diversity profile and empirical profile; the difference in diversity between any two assemblages can be evaluated and tested for significance.

Step 3. Non-asymptotic coverage-based rarefaction and extrapolation analysis for orders $q = 0$, 1 and 2

When sampling data do not contain sufficient information to accurately infer true diversity, fair comparisons of diversity across multiple assemblages should be made by standardizing the sample coverage (i.e., comparing diversity for a standardized fraction of an assemblage's individuals). This comparison can be performed based on seamless integration of coverage-based rarefaction and extrapolation sampling curves up to a maximum coverage value of $C_{\max}$ defined in Section 4.4 (i.e., the level of coverage reached by the sample that attains the lowest coverage when all samples are extrapolated to double the reference sample size).

Step 4. An evenness profile

As discussed in Section 4.5, the magnitude of the normalized slopes of the diversity profile can be used to derive measures of evenness among species abundances. We suggest assessing and comparing evenness at the coverage value of $C_{\max}$ among the samples compared. In our analysis, we provide the evenness profile that depicts the evenness measure $^qE = (^qD - 1)/(S - 1)$ with respect to order q, where diversity ^qD and S are computed at the coverage value of $C_{\max}$. We also provide the widely used Pielou's (1966) J' index, which is also computed at the same coverage value. All these evenness measures are standardized to the range of [0, 1] to adjust for the effect of differing species richness.

The proposed estimated sample completeness profile, asymptotic diversity estimates and non-asymptotic rarefaction and extrapolation sampling curves along with the associated evenness profile can be computed from the online freeware application iNEXT-4steps available from http://chao.stat.nthu.edu.tw/wordpress/software_download/. This freeware is an expanded and updated version of iNEXT (Hsieh, Ma, & Chao, 2016).

6 | EMPIRICAL EXAMPLES

In this section, two abundance-based datasets (fossil ostracods and spiders) and two incidence-based datasets (woody plants and stony corals) are analyzed to illustrate our suggested four-step procedure. We provide step-by-step analysis and interpretation for Example 1. Parallel analysis and interpretation can be applied to the other three examples, but details are omitted to avoid repetitive text.

6.1 | Example 1 (abundance-based fossil data)

Shin et al. (2019) investigated species richness patterns in Neogene fossil marine ostracods from Java, Indonesia. In their data, a total of 171 species were identified from nine sediment samples from three different geological ages: Middle Miocene (1 sample), Late Miocene (3 samples) and Pliocene (5 samples). They concluded that there was a significant increase in species richness from the Late Miocene to the Pliocene. Their comparisons were based on the conventional E50 rarefaction (Shin et al., 2019). That is, the expected number of species for a rarefied sample size of 50 is first calculated for each sample; these expected values are then compared across samples. For illustration purposes, we pool the data of Middle Miocene and Late Miocene and compare the pooled data (referred to as "Miocene," four samples pooled) with those of the Pliocene (five samples). In the Miocene samples, 26 species were represented by 306 specimens, whereas in the Pliocene samples, 167 species were represented by 4,177 specimens. The species frequency data were summarized in the frequency counts and are tabulated in Table S4.1.

Based on the conventional E50 rarefaction, the estimated species richness for the Miocene is 13.6, whereas the corresponding value for the Pliocene is 21.3; the difference is about 21.3–13.6 = 7.7, and the species richness of the Pliocene is 1.6 times higher than that of the Miocene. However, in the conventional E50 rarefaction, the sample coverage for the rarefied sample of size 50 for the Miocene is estimated to be 88.3% whereas the corresponding estimate for the Pliocene is 73.8% (based on iNEXT output, not shown), implying that the conventional comparison for a low rarefied size of 50 is not based on samples with equal coverage and thus the difference in species richness between the two assemblages is much compressed (Chao & Jost, 2012).

Here, we demonstrate that our methodologies can provide additional information and insights into the assessment of the Neogene marine ostracod diversity. For each dataset, Figure 1a–e shows, respectively, the estimated sample completeness profiles, size-based rarefaction and extrapolation curves, asymptotic and empirical diversity profiles, coverage-based rarefaction and extrapolation curves and the evenness profile. Numerical values for the three orders $q = 0$, 1 and 2 are provided in Table 2 (left half). More details on the uncertainties for the asymptotic diversity estimates and the associated 95% confidence intervals are provided in Table S4.1. As suggested in the preceding section, we analyze the data in the following four steps:

Step 1. Assessment of sample completeness profiles (Table 2 and Figure 1a)

Figure 1a shows that the two estimated sample completeness profiles are both increasing with diversity order, implying that there was undetected diversity within each dataset. The two profiles cross and are statistically indistinguishable due to the sparse Miocene data, which cause wide confidence bands. Although the two reference sample sizes differ greatly, the two estimated sample-completeness profiles in Figure 1a are generally close to each other. For example, the sample coverage value for the Miocene data (306 specimens) is 97.7% and the corresponding value for the Pliocene data (4,177 specimens) is 98.8%. Detailed interpretations are given below.

- The estimated sample completeness for $q = 0$, 1 and 2 for the Miocene data are, respectively, 81.0%, 97.7% and 99.9%. This means that the data cover at most 81% of the total species in the assemblage; the detected species cover about 97.7% of the assemblage's individuals, and about 99.9% of the individuals if the focus is on highly abundant species. In other words, the undetected proportion of species is at least 19% of the total species; the undetected species cover about 2.3% of the assemblage's individuals, or about 0.1% of the individuals if only highly abundant species are considered.
- The estimated sample completeness of $q = 0$, 1 and 2 for the Pliocene data are, respectively, 79.0%, 98.8% and 99.9%. Similar interpretations can be made, analogous to those for the Miocene data.

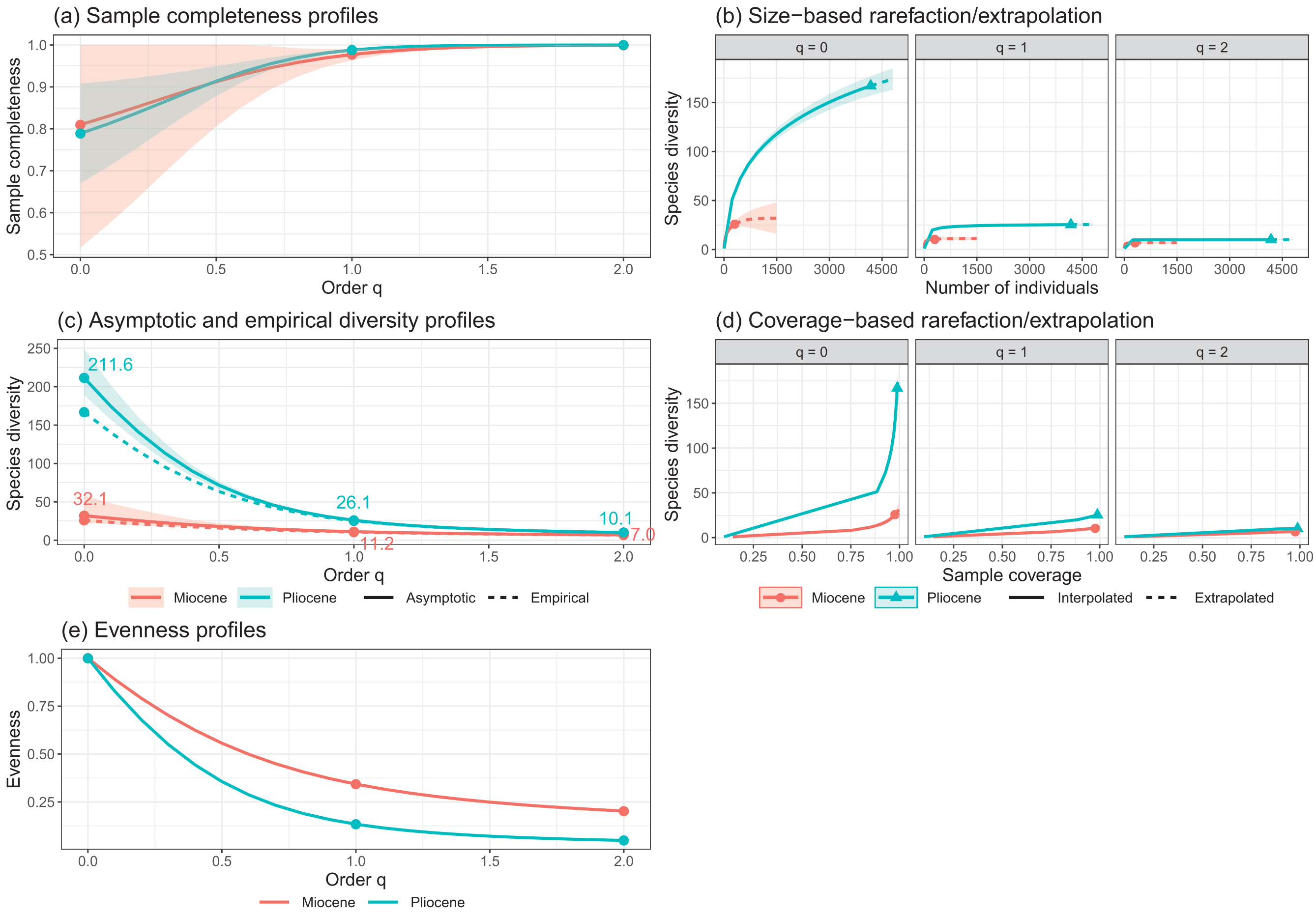

FIGURE 1 (a) The plots of estimated sample completeness curves as a function of order q between 0 and 2 in the Miocene data ($S_{\text{obs}} = 26$, $n = 306$) and Pliocene data ($S_{\text{obs}} = 167$, $n = 4{,}177$) for fossil marine ostracods from Java, Indonesia (Shin et al., 2019). (b) Sample-size-based rarefaction (solid lines) and extrapolation curves (dashed lines) up to size 1,500 for Miocene data and 4,800 for Pliocene data to allow better visualization. (b) The asymptotic estimates of diversity profiles (solid lines) and empirical diversity profiles (dotted lines); numerical values refer to the estimated asymptotic diversities. (d) Coverage-based rarefaction (solid lines) and extrapolation (dashed lines) curves up to the corresponding coverage value for size 1,500 (for Miocene data) and 4,800 (for Pliocene data). (e) Evenness profile as a function of order q, for $0 < q \leq 2$, based on the normalized slope of Hill numbers. Solid dots and triangles denote observed data points. All shaded areas in (a)–(d) denote 95% confidence bands obtained from a bootstrap method with 100 replications. Some bands are invisible due to narrow widths. Numerical values for the three special cases of $q = 0$, 1 and 2 are shown in Table 2 (left half) [Color figure can be viewed at wileyonlinelibrary.com]

Step 2. Size-based rarefaction and extrapolation analysis and the asymptotic diversity profiles (Table 2 and Figure 1b,c)

(*Step 2a*) Figure 1b reveals that, for each dataset, the size-based rarefaction and extrapolation sampling curves for diversity of orders $q = 1$ and $q = 2$ stabilize, implying that our asymptotic diversity estimates for these two measures work satisfactorily to infer true diversities. However, neither of the sampling curves for species richness ($q = 0$), extrapolated up to double the reference sample size, stays at a fixed level, suggesting that the current data do not contain sufficient information to accurately estimate true species richness within each assemblage; our asymptotic estimate of species richness (the Chao1 estimate) thus represents a minimum species richness.

(*Step 2b*) Comparing the estimated asymptotic diversity profile (solid lines in Figure 1c) and the corresponding observed/empirical diversity profile (dashed lines in Figure 1c), we can assess the extent of undetected diversity within each dataset and the diversity difference between the two datasets, separately for $q = 0$, 1 and 2 (Table 2). Details are elaborated below:

- The undetected species richness within the Miocene and Pliocene are, respectively, at least 6.1 (≥ 19%) and 44.6 (≥ 21%). Since these estimates are lower bounds, as indicated in Step 2a, the degree of difference in true

TABLE 2 The numerical values for the three special cases of q = 0, 1 and 2 for abundance-based fossil marine ostracods collected from Java, Indonesia (Example 1, left half, Shin et al., 2019) and spider data collected from the Bavarian Forest National Park, Germany (Example 2, right half, Thorn et al., 2017)

Example 1 (Fossil data, Figure 1a–e)				**Example 2 (Spider data, Figure 2a–e)**			
Step 1. Sample completeness profiles (panel a in each figure)							
Completeness	$q = 0$	$q = 1$	$q = 2$	Completeness	$q = 0$	$q = 1$	$q = 2$
Miocene	81.0%	97.7%	99.9%	Open forest	76.3%	98.6%	99.9%
Pliocene	79.0%	98.8%	99.9%	Closed forest	60.3%	98.9%	99.9%
Step 2. Asymptotic analysis (panels b and c in each figure)							
Diversity	$q = 0$	$q = 1$	$q = 2$		$q = 0$	$q = 1$	$q = 2$
Miocene				Open forest			
Asymptotic	32.1	11.2	7.0	Asymptotic	96.3	16.8	9.5
Empirical	26	10.6	6.8	Empirical	74	16.3	9.4
Undetected	6.1	0.6	0.2	Undetected	22.3	0.5	0.1
Pliocene				Closed forest			
Asymptotic	211.6	26.1	10.1	Asymptotic	72.1	10.3	5.7
Empirical	167	25.3	10.1	Empirical	44	10.0	5.7
Undetected	44.6	0.8	0.0	Undetected	28.1	0.3	0.0
Step 3. Non-asymptotic coverage-based rarefaction and extrapolation (panel d in each figure)							
Maximum standardized coverage $C_{\max} = 99.3\%$				Maximum standardized coverage $C_{\max} = 99.4\%$			
Diversity	$q = 0$	$q = 1$	$q = 2$	Diversity	$q = 0$	$q = 1$	$q = 2$
Miocene	30.2	11.0	6.9	Open forest	86.9	16.6	9.4
Pliocene	185.5	25.6	10.0	Closed forest	56.2	10.2	5.7
Step 4: Evenness among species abundances (panel e in each figure)							
Evenness	Pielou J'	$q = 1$	$q = 2$	Evenness	Pielou J'	$q = 1$	$q = 2$
Miocene	0.70	0.34	0.20	Open forest	0.63	0.18	0.10
Pliocene	0.62	0.13	0.05	Closed forest	0.58	0.17	0.09

Note: See Figures 1 and 2 for the corresponding profiles with 95% confidence intervals.

species richness of the entire assemblages cannot be precisely assessed.

- The undetected Shannon diversity within the Miocene and Pliocene are, respectively, 0.6 and 0.8. That is, only about one abundant species was not detected within each geological age (Table 2). As explained in Step 2a, these asymptotic values represent accurate estimates of the true diversities, and the difference between the two assemblages, with respect to abundant species, is 26.1–11.2 = 14.9; the difference is statistically significant because the two 95% confidence bands in Figure 1c do not overlap.
- The undetected Simpson diversity within the Miocene and Pliocene are, respectively, 0.2 and 0, implying that nearly all highly abundant species were detected. As explained in Step 2a, these asymptotic values represent accurate estimates of the true diversities, and the difference between the two assemblages with respect to highly abundant species is 10.1–7.0 = 3.1; this difference is also statistically significant.

Step 3. Non-asymptotic analysis for diversity orders q = 0, 1 and 2 (Table 2 and Figure 1d)

The above asymptotic analysis implies that, for diversities of q = 1 and 2, the true diversity of the entire assemblage of the Pliocene is significantly higher than that of the Miocene. Figure 1d reveals that this conclusion is valid not only for comparing the two entire assemblages, but also for any standardized sample coverage up to unity. For species richness, although our data are insufficient to infer the true richness of the entire assemblage, diversity and evenness measures can be computed up to a standardized coverage value of $C_{\max}$ = 99.3% (the lower coverage of the two extrapolated samples when each is extrapolated to double the reference sample size), as previously defined. A tiny fraction of an assemblage's individuals may contain infinitely many species because of the potential presence of vanishingly rare species. Here, the richness of the remaining 0.7% of the assemblage is not estimable from the data, due to insufficient information for the rarest 0.7% of the individuals. In this case, coverage-based sampling curves enable us to make sensible inferences and fair comparisons of diversity profiles and their slopes for any standardized assemblage fraction up to 99.3%.

For the maximum standardized coverage value of 99.3%, the corresponding richness estimate for the Miocene is 30.2, whereas for the Pliocene it is 185.5 (see Table 2). The degree of difference between the two stages can be precisely assessed: the difference in species richness between the Pliocene and Miocene is 155.3; the Pliocene assemblage is 6.1 times higher than the Miocene assemblage. Note that our estimated difference and estimated ratio, based on a standardized coverage of 99.3%, are much higher than those obtained from a standardized sample of size 50 (a difference of 7.7 and a ratio of 1.6). The diversity value of $q = 1$ for a 99.3% assemblage fraction differs very little from that of the entire assemblage. The same conclusion is valid for the diversity of order $q = 2$.

Step 4. Evenness profile

Figure 1e shows the evenness profile for diversity orders $0 < q \leq 2$. For all values of q, the evenness profile and Pielou's measure (Table 2) are computed for a coverage value of 99.3%. All measures consistently show that the evenness among species abundances in the Miocene is higher than that in the Pliocene for a standardized 99.3% assemblage fraction.

In summary, in the conventional E50 comparison, most data are discarded and the difference in species richness between the Miocene and Pliocene is much compressed.

1. Our sample completeness analysis shows that the undetected species richness within the Miocene data and Pliocene data are, respectively, at least 6.1 ($\geq$ 19%) and 44.6 ($\geq$ 21%); within each geological age, only about one abundant species was not detected and nearly all highly abundant species were detected. The undetected species in the Miocene assemblage cover about 2.3% of the assemblage's individuals, and about 1.2% for the Pliocene assemblage.
2. For Shannon diversity ($q = 1$) and Simpson diversity ($q = 2$), our analysis demonstrates that the Pliocene assemblage is significantly more diverse than the Miocene assemblage; this conclusion is valid for any standardized coverage up to 100%, but for species richness ($q = 0$), it is valid up to a 99.3% assemblage fraction.
3. Under the maximum coverage of 99.3%, the difference in species richness, Shannon diversity and Simpson diversity between the Pliocene and Miocene assemblages are, respectively, 155.3, 14.6 and 3.1 species (Table 2, Step 3). Thus, the major difference between the two ages lies in rare species. That is, according to this fossil ostracod dataset, the increase in richness from the Miocene to the Pliocene is mainly due to an increase in rare species. Our analysis can clearly quantify the magnitude of increase and test the significance in diversity difference, as demonstrated above.

6.2 | Example 2 (abundance-based spider data)

These data were sampled in a mountain forest ecosystem in the Bavarian Forest National Park, Germany (Thorn et al., 2016, 2017). A total of 12 experimental plots were established in *closed forest* stands (six plots) and *open forest* stands with naturally occurring gaps and edges (six plots) to assess the effects of microclimate on communities of epigeal (ground-dwelling) spiders. Epigeal spiders were sampled over 3 years with four pitfall traps in each plot, yielding a total of 3,171 individuals belonging to 85 species recorded in the pooled habitat. In the open forest, there were 1,760 individuals representing 74 species, whereas in the closed forest, there were 1,411 individuals representing 44 species. The species frequency data are summarized into frequency counts and are tabulated in Table S4.2.

Figure 2a–e shows, respectively, the estimated sample completeness profiles, size-based rarefaction and extrapolation curves, asymptotic and empirical diversity profiles, coverage-based rarefaction and extrapolation curves, and the evenness profile for the spider dataset. All the numerical values, specifically for the three orders $q = 0$, 1 and 2, are provided in Table 2 (right half). More details on the uncertainties for the asymptotic diversity estimates and the associated 95% confidence intervals are provided in Table S4.2. The four steps for assessing sample completeness and comparing diversity between the two forests are generally parallel to those in Example 1. Readers may refer to Example 1 for interpretation. We thus omit most details and only summarize the conclusions below:

1. Figure 2a shows that the estimated sample completeness for the open forest is higher than that of the closed forest when $q < 0.5$, although confidence intervals overlap. For $q > 0.5$, the sample completeness of the two forests is similar. This result suggests that a higher proportion of rare species in the closed forest were not detected. For $q = 0$, the estimated sample completeness for the open forest and the

TABLE 3 The numerical values for the three special cases of $q = 0$, 1 and 2 for incidence-based woody plant data collected in Taiwan (Example 3, left half, Chiou et al., 2009, Li et al., 2013) and for stony coral data in the Coral Triangle and Madagascar area (Example 4, right half)

Example 3 (Woody plant data, Figure 3a–e)				**Example 4 (Stony coral data, Figure 5a–e)**			
Step 1. Sample completeness profiles (panel a in each figure)							
Completeness	$q = 0$	$q = 1$	$q = 2$	Completeness	$q = 0$	$q = 1$	$q = 2$
Monsoon	78.0%	98.9%	99.9%	Madagascar	73.7%	99.5%	99.9%
Upper cloud	77.6%	98.2%	99.9%	Coral triangle	91.5%	99.1%	99.9%
Step 2. Asymptotic analysis (panels b and c in each figure)							
Diversity	$q = 0$	$q = 1$	$q = 2$	Diversity	$q = 0$	$q = 1$	$q = 2$
Monsoon				Madagascar			
Asymptotic	421.7	150.2	103.3	Asymptotic	592.3	360.1	331.3
Empirical	329	145.6	102.3	Empirical	437	350.9	322.3
Undetected	92.7	4.6	1.0	Undetected	155.3	9.2	9.0
Upper cloud				Coral triangle			
Asymptotic	307.8	110.5	72.2	Asymptotic	421.1	203.0	132.8
Empirical	239	105.5	71.2	Empirical	386	195.6	130.0
Undetected	68.8	5.0	1.0	Undetected	35.1	7.4	2.8
Step 3. Non-asymptotic coverage-based rarefaction and extrapolation (panel d in each figure)							
Maximum standardized coverage $C_{\max} = 99.3\%$				Maximum standardized coverage $C_{\max} = 99.6\%$			
Diversity	$q = 0$	$q = 1$	$q = 2$	Diversity	$q = 0$	$q = 1$	$q = 2$
Monsoon	363.3	147.5	102.7	Madagascar	474.0	355.4	326.2
Upper cloud	280.6	108.6	71.7	Coral triangle	406.0	198.7	131.0
Step 4. Evenness among species abundances (panel e in each figure)							
Evenness	Pielou J'	$q = 1$	$q = 2$	Evenness	Pielou J'	$q = 1$	$q = 2$
Monsoon	0.85	0.40	0.28	Madagascar	0.95	0.75	0.69
Upper cloud	0.83	0.38	0.25	Coral triangle	0.88	0.49	0.32

Note: Data for Example 4 were taken from two databases: the Global Biodiversity Information Facility (GBIF) and the Ocean Biogeographic Information System (OBIS). See Figures 3 and 5 for the corresponding profiles with 95% confidence intervals.

closed forest are, respectively, 76.3% and 60.3%; the corresponding values for $q = 1$ are 98.6% and 98.9%. It then follows from Table 2 that at least 1–76.3% = 23.7% (≥ 22.3 species) of the species within the open-forest assemblage and at least 1–60.3% = 39.7% (≥ 28.1 species) of the species within the closed-forest assemblage were not detected in the sample. The undetected species in the open forest cover about 1–98.6% = 1.4% of the assemblage's individuals, and about 1–98.9% = 1.1% for the closed forest. Based on the estimated undetected diversities of $q = 1$ and $q = 2$ in Table 2, nearly all abundant species and highly abundant species had been found, within the data for each forest type.

2. Figure 2b,c shows that, for Shannon diversity ($q = 1$) and Simpson diversity ($q = 2$), the open forest is significantly more diverse than the closed forest for any assemblage fraction up to unity. Based on the asymptotic results (Table 2), there is a moderate difference (~6.5) between the two entire assemblages with respect to abundant species, and a small difference (~3.8) with respect to highly abundant species.
3. For species richness, although our data are insufficient to infer the true richness of the entire assemblage, inference and significance testing can be performed up to a standardized coverage value of $C_{\max} = 99.4\%$ (the minimum coverage of two samples extrapolated to double the size of the reference sample). Under a standardized coverage of 99.4%, the difference in species richness, Shannon diversity and Simpson diversity between the open and closed forests are, respectively, 30.7, 6.4 and 3.7 species (Figure 2d, Table 2). All differences are significant, as shown by the non-overlapping confidence intervals in Figure 2d ($q = 0$).
4. Under the coverage value of 99.4%, Pielou's evenness measure (Table 2) shows that the evenness among species abundances in the open forest is higher than

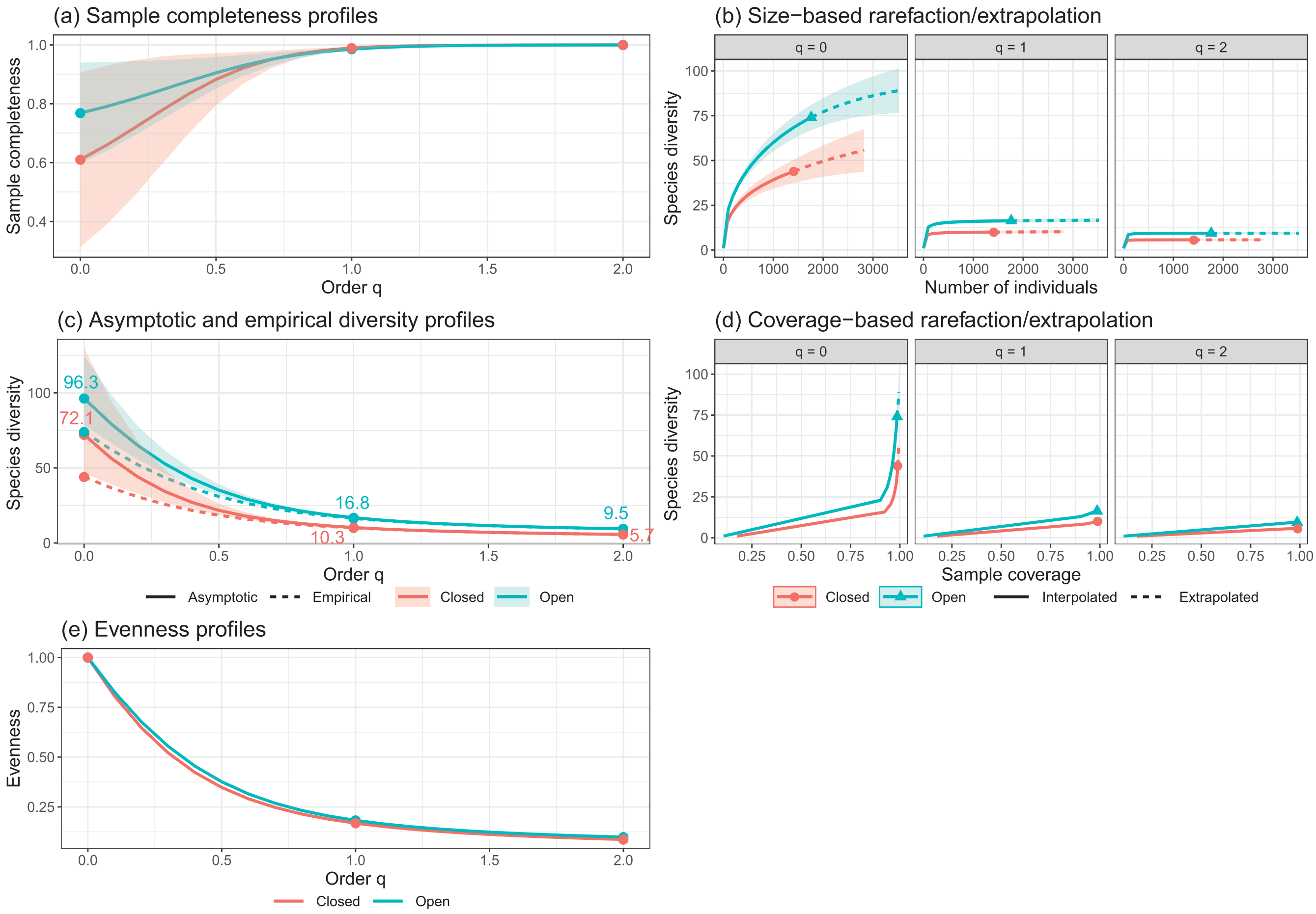

FIGURE 2 (a) The plots of estimated sample completeness curves as a function of order q between 0 and 2 for spider species data collected in a closed forest ($S_{\text{obs}} = 44$, $n = 1{,}411$) and an open forest ($S_{\text{obs}} = 74$, $n = 1{,}760$) in the Bavarian Forest National Park, Germany (Thorn et al., 2016; Thorn, Bässler, Svoboda, & Müller, 2017) (b) Size-based rarefaction (solid lines) and extrapolation (dashed lines) curves up to double the reference sample size. (c) The asymptotic estimates of diversity profiles (solid lines) and empirical diversity profiles (dotted lines); numerical values refer to the estimated asymptotic diversities. (d) Coverage-based rarefaction (solid lines) and extrapolation (dashed lines) curves up to the corresponding coverage value for a doubling of each reference sample size. (e) Evenness profile as a function of order q, $0 < q \leq 2$, based on the normalized slope of Hill numbers. Solid dots and triangles denote observed data points. All shaded areas in (a)–(d) denote 95% confidence bands obtained from a bootstrap method with 100 replications. Some bands are invisible due to narrow widths. Numerical values for the three special cases of $q = 0$, 1 and 2 are shown in Table 2 (right half) [Color figure can be viewed at wileyonlinelibrary.com]

that in the closed forest. The profile (Figure 2e), however, reveals that the evenness values for the two forests are very close for any order q between 0 and 2.

6.3 | Example 3 (incidence-based woody plant data)

The datasets considered here are a subset of the National Vegetation Database of Taiwan (AS-TW-001), sampled between 2003 and 2007 as part of the first national vegetation inventory project (Chiou et al., 2009). Over 3,600 vegetation plots, each 20 m × 20 m in size, were set up in various locations in Taiwan, and all woody plant individuals taller than 2 m were recorded in each plot. For illustration here, we selected only plots belonging to two vegetation types (according to Li et al., 2013): *Pyrenaria-Machilus* subtropical winter monsoon forest and *Chamaecyparis* montane mixed cloud forest, sampled in the northern part of Taiwan (in ecoregions 7 and 8 according to Su, 1985). Because spatial clustering prevails in woody plants, individual plants cannot be regarded as independent sampling units, violating the basic sampling assumptions for the model based on abundance data. We thus use incidence data to avoid this violation. Plots within each vegetation type were aggregated into the incidence-based datasets and are referred to as the *monsoon forest* and *upper cloud forest* vegetation types in the following analysis.

In the monsoon forest, 329 species and 6,814 incidences were recorded in 191 plots. In the upper cloud

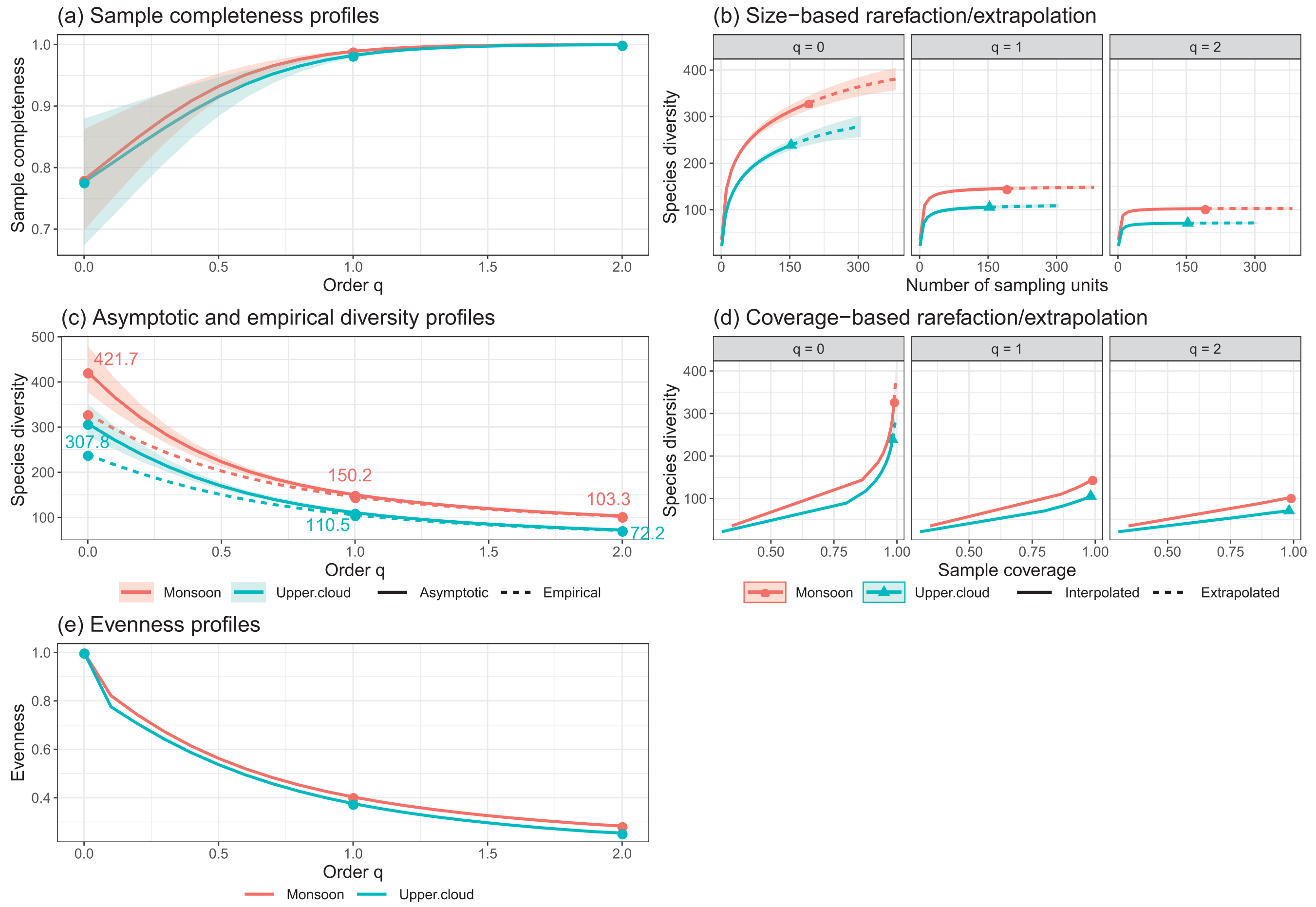

FIGURE 3 (a) The plots of estimated sample completeness curves as a function of order q between 0 and 2 for woody plant data in the subtropical winter monsoon forest (called *monsoon forest* here, $S_{\text{obs}} = 329$, $T = 191$ plots) and montane mixed cloud forest (*upper cloud forest*, $S_{\text{obs}} = 239$, $T = 153$ plots) vegetation types in Taiwan (The National Vegetation Database of Taiwan, Chiou et al., 2009). (b) Size-based rarefaction (solid lines) and extrapolation (dashed lines) curves up to double the reference sample size. (c) The asymptotic estimates of diversity profiles (solid lines) and empirical diversity profiles (dotted lines); numerical values refer to the estimated asymptotic diversities. (d) Coverage-based rarefaction (solid lines) and extrapolation (dashed lines) curves up to the corresponding coverage value for a doubling of each reference sample size. (e) Evenness profile as a function of order q, $0 < q \leq 2$, based on the normalized slope of Hill numbers. Solid dots and triangles denote observed data points. All shaded areas in (a)–(d) denote 95% confidence bands obtained from a bootstrap method with 100 replications. Some bands are invisible due to narrow widths. Numerical values for the three special cases of $q = 0$, 1 and 2 are shown in Table 3 (left half) [Color figure can be viewed at wileyonlinelibrary.com]

forest, 239 species and 3,371 incidences were recorded in 153 plots (each plot is regarded as a sampling unit). The incidence-based species frequency data are summarized as incidence frequency counts and are tabulated in Table S4.3. For each dataset, Figure 3a–e shows, respectively, the estimated sample completeness profiles, size-based rarefaction and extrapolation curves, asymptotic and empirical diversity profiles, coverage-based rarefaction and extrapolation curves, and the evenness profile. All numerical values specifically for the three orders $q = 0$, 1 and 2 are provided in Table 3 (left half). More details on the uncertainties and 95% confidence intervals for the asymptotic diversity estimates are provided in Table S4.3. The four steps for assessing sample completeness and comparing diversity between the two forests are generally parallel to those in Example 1; here we only summarize some major conclusions below:

1. Figure 3a reveals that although the sample completeness profile for the monsoon forest is slightly higher than the profile for the upper cloud forest, the two estimated curves stay very close, with their corresponding two confidence bands completely overlapping. For $q = 0$, the estimated sample completeness for the monsoon forest and the upper cloud forest are, respectively, 78% and 77.6%; the corresponding values for $q = 1$ are 98.9% and 98.2%. Table 3 then shows that at least 1–78% = 22% (≥ 92.7 species) of the species

within the monsoon forest and at least $1–77.6\% = 22.4\%$ (≥ 68.8 species) of the species within the upper cloud forest were not detected in the samples. The undetected species cover about $1–98.9\% = 1.1\%$ of the assemblage's *incidences* for the monsoon forest, and about $1–98.2\% = 1.8\%$ for the upper cloud forest. Based on the estimated undetected diversities of $q = 1$ and $q = 2$ in Table 3, about 4–5 frequent species and about 1 highly frequent species were not detected within the data of each vegetation type.

2. Figure 3b,c shows that for diversity of orders $q = 1$ and 2, the monsoon forest is significantly more diverse than the upper cloud forest for any fraction up to entire assemblages. Based on the asymptotic results, the difference between the two forest types is 39.7 with respect to frequent species (for $q = 1$), whereas the difference is about 31.1 with respect to the highly frequent species (for $q = 2$).
3. For species richness, although our data are insufficient to infer the true richness of the entire assemblage, inference and significance testing can be made up to a standardized coverage value of $C_{\text{max}} = 99.3\%$. Under the maximum assemblage fraction, Figure 3d and Table 3 show that the major difference between the two forest types lies in infrequent species, as reflected by the differences of 82.7 in species richness ($q = 0$), 38.9 for frequent species ($q = 1$) and 31.0 for highly frequent species ($q = 2$). All differences are significant.
4. Under the coverage value of 99.3%, the evenness profile and Pielou's measure (Figure 3e and Table 3) consistently show that the evenness among species occurrences in the monsoon forest is slightly higher than that in the upper cloud forest.

6.4 | Example 4 (incidence-based tropical stony coral data)

The data for Scleractinia species considered in this example were taken from two databases: the Global Biodiversity Information Facility (GBIF; https://www.gbif.org/) and the Ocean Biogeographic Information System (OBIS; http://www.iobis.org/). Kusumoto et al. (2020) provided a more detailed description of how this dataset was compiled and how coral taxonomy was scrutinized. The global area was first divided into grids of 5 × 5°. We selected 30 such grids covering the Coral Triangle (Figure 4), which is empirically regarded as a diversity hot spot of corals and other marine organisms (e.g., Asaad, Lundquist, Erdmann, & Costello, 2018; Briggs, 2003; Veron et al., 2009). We also selected the same number of grids covering the Madagascar area (Figure 4) where another diversity peak was detected by a global-scale analysis by Kusumoto et al. (2020).

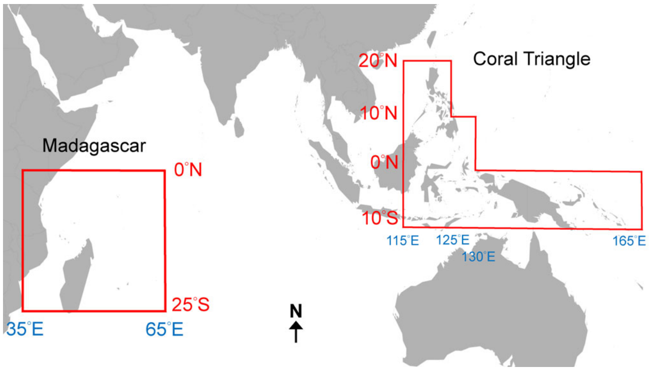

FIGURE 4 The Coral Triangle and the Madagascar area: there are 30 grids of 5 × 5° in each area. Global-scale sampling locations for tropical stony corals (Scleractinia) used in the analysis are detailed in Kusumoto et al. (2020) [Color figure can be viewed at wileyonlinelibrary.com]

Each defined area was further divided into sub-grids of 0.01 × 0.01°, and species incidence data were recorded in each sub-grid. Here, a sub-grid of size 0.01 × 0.01° is regarded as a sampling unit in the incidence-data framework. In the Coral Triangle, 386 species and 5,926 incidences were recorded in 665 sub-grids. In the Madagascar area, 437 species and 10,079 incidences were recorded in 198 sub-grids. The incidence-based species frequency data were summarized as incidence frequency counts (Table S4.4). Dataset details are provided in the supplement of Kusumoto et al. (2020).

For each dataset, Figure 5a–e shows, respectively, the estimated sample completeness profiles, size-based rarefaction and extrapolation curves, asymptotic and empirical diversity profiles, coverage-based rarefaction and extrapolation curves, and the evenness profile for the spider dataset. All numerical values mentioned in the following analysis are provided in Table 3 (right half). More details on the uncertainties for the asymptotic diversity estimates and the associated 95% confidence intervals appear in Table S4.4. The four steps for assessing sample completeness and comparing diversity between the two forests are generally parallel to those in Example 1; here we only summarize some major conclusions below:

1. The increasing pattern of the sample completeness profiles in Figure 5a for each area implies that there are undetected species within each dataset. For any fixed order $q < 0.5$, sample completeness for the Madagascar area is much lower than that of the Coral Triangle. This pattern suggests that a higher proportion of rare species were overlooked in the

Madagascar area than in the Coral Triangle. For any fixed order $q \geq 0.5$, the two areas are very close in sample completeness. For $q = 0$, the estimated sample completeness for the Madagascar area and the Coral Triangle are, respectively, 73.7% and 91.5%; the corresponding values for $q = 1$ are 99.5% and 99.1%. It then follows from Table 3 that at least 1–73.7% = 26.3% (≥ 155 species) of the species in the Madagascar area and at least 1–91.5% = 8.5% (≥ 35 species) of the species in the Coral Triangle were not detected. The undetected species cover about 1–99.5% = 0.5% of incidences in the Madagascar area, and about 0.9% in the Coral Triangle. Based on the estimated undetected diversities of $q = 1$ and $q = 2$ in Table 3, about 9.2 and 7.4 frequent species were not detected in the Madagascar area and in the Coral Triangle, respectively; the number of undetected highly frequent species for the two areas were, respectively, about 9.0 and 2.8.

2. Figure 5b,c shows that for each diversity order $q = 1$ and 2, the Madagascar area is significantly more diverse than the Coral Triangle for any fraction up to entire assemblages. The asymptotic estimates for entire assemblages reveal that the difference with respect to frequent species is about 157.1, and is about 198.5 with respect to highly frequent species.
3. For species richness, although our data are insufficient to infer the true richness of the entire

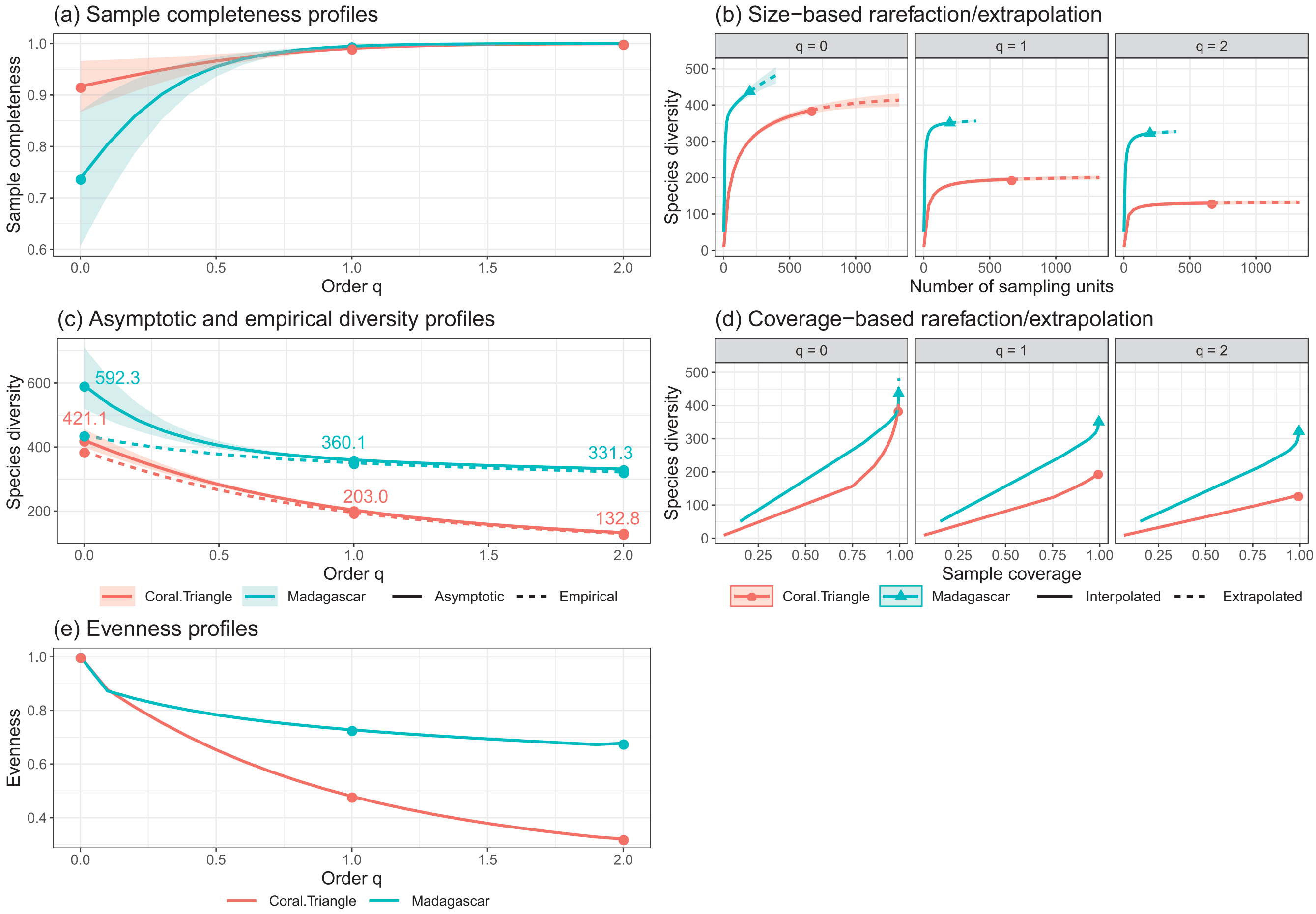

FIGURE 5 (a) The plots of estimated sample completeness curves as a function of order q between 0 and 2 for stony coral data in the Coral Triangle ($S_{\text{obs}} = 386$, $T = 665$ sub-grids with at least one occurrence) and Madagascar area ($S_{\text{obs}} = 437$, $T = 198$ sub-grids with at least one occurrence). Data were taken from two databases: the Global Biodiversity Information Facility (GBIF) and the Ocean Biogeographic Information System (OBIS). (b) Size-based rarefaction (solid lines) and extrapolation (dashed lines) curves up to double the reference sample size. (c) The asymptotic estimates of diversity profiles (solid lines) and empirical diversity profiles (dotted lines); numerical values refer to the estimated asymptotic diversities. (d) Coverage-based rarefaction (solid lines) and extrapolation (dashed lines) curves up to the corresponding coverage value for a doubling of each reference sample size. (e) Evenness profile as a function of order q, $0 < q \leq 2$, based on the normalized slope of Hill numbers. Solid dots and triangles denote observed data points. All shaded areas in (a)–(d) denote 95% confidence bands obtained from a bootstrap method with 100 replications. Some bands are invisible due to narrow widths. Numerical values for the three special cases of $q = 0$, 1 and 2 are shown in Table 3 (right half) [Color figure can be viewed at wileyonlinelibrary.com]

assemblage, inference and significance testing can be performed up to a standardized coverage value of $C_{\max} = 99.6\%$. Under a standardized coverage of 99.6%, the estimated richness estimate is 474.0 for the Madagascar area, and 406.0 for the triangle. The difference in species richness between the Madagascar area and Coral Triangle is 68, and the difference is significant. For a 99.6% assemblage fraction, the difference in Shannon diversity ($q = 1$) between the two areas is about 156.7 and about 195.2 for Simpson diversity ($q = 2$); these differ very little from those of the entire assemblages.

4. Under the coverage value of 99.6%, the evenness profile and Pielou's measure (Figure 5e and Table 3) show that evenness among species occurrences in the Madagascar area is much higher than that in the Triangle area.

It should be noted, however, that our diversity comparison between the two areas is not consistent with previous studies which have suggested that the highest peak of species richness lies within the Coral Triangle (Veron et al., 2009); see Section 7.1 for Discussion.

7 | DISCUSSION

All of our examples exhibit the following consistent patterns in the four recommended analysis steps. Whether the same patterns are valid for other datasets from highly-diverse assemblages requires further investigation. All the following panel labels refer to those in Figures 1–3 and 5.

Step 1. Assessment of sample completeness profiles (Panel a)

All the estimated profiles in the examples increase with order q, revealing that there are undetected species in each dataset. When $q > 2$, nearly all profiles approach the level of unity, indicating that all highly abundant species (for abundance data) or highly frequent species (for incidence data) had been detected in the reference samples.

Step 2. Size-based rarefaction and extrapolation and asymptotic diversity profile (Panels b and c)

All size-based rarefaction and extrapolation sampling curves for diversity measures of $q = 1$ and 2 level off, implying that our asymptotic Shannon and Simpson diversity values can be used to accurately infer true diversities of entire assemblages, and the difference in these two diversities among assemblages can be evaluated and tested for significance in difference. However, none of the curves for species richness ($q = 0$) stabilize, up to double the size of the reference sample, implying that our asymptotic species richness estimates represent only lower bounds.

Step 3. Non-asymptotic coverage-based rarefaction and extrapolation for diversity orders $q = 0$, 1 and 2 (Panel d)

The difference in species richness among assemblages for all examples were evaluated and tested for a standardized fraction of assemblages up to a maximum based on the coverage-based rarefaction and extrapolation sampling curves. The maximum coverage values in the four examples were all > 99%. As indicated in Example 1, due to the potential presence of vanishingly rare species, a tiny assemblage fraction may contain infinitely many species, making species richness in that fraction non-estimable.

Step 4. Evenness profile for q between 0 and 2 (Panel e)

For a specified maximum standardized fraction, not only species diversity but also evenness profiles can be estimated and compared across assemblages.

7.1 | Our findings in the stony coral example

Our diversity analysis and comparison in the first three examples generally conform to the results obtained from previous studies (Li et al., 2013; Shin et al., 2019; Thorn et al., 2016, 2017). However, in Example 4 (stony coral datasets), our findings are not consistent with previous studies that have suggested that the highest peak of species richness lies within the Coral Triangle (Veron et al., 2009). This inconsistency may be due to a difference in data sources and quality. The Coral Triangle was defined by stacking the expert range maps of coral species (Veron & Stafford-Smith, 2000), whereas our analysis was based on the occurrence data from two large data bases. Note that for the observed species richness in the two databases for the two areas (defined in Figure 4), the Madagascar area has more species (437 species) than the Coral Triangle (386 species). Thus, it is not unexpected that all our analysis procedures reveal that the Madagascar area is more diverse than the Coral Triangle. The Coral Triangle and the Madagascar area are of the same size in area; stony corals occurred in 665 sub-grids in the Coral Triangle but in only 198 sub-grids in the Madagascar area. As also revealed from our sample completeness profile (Figure 5a) and asymptotic analysis, it is likely that the Madagascar area remains severely undersampled and a large fraction of rare species has not yet been detected. Kusumoto et al. (2020) conducted a biodiversity estimation analysis at global scale using occurrence records and showed that the species richness of stony corals increased near edges, rather than the center, of the Coral Triangle. The Coastal Indo-Pacific and

Offshore Indian Ocean realms (of species endemicity) extend to Madagascar where they meet the Southern African realm, also known for high endemicity (Costello et al., 2017). Further research should confirm the present findings, which may be due to the Madagascar area overlapping with three marine realms. Our results could help guide the effective allocation of future sampling efforts and address coral biodiversity information shortfalls.

7.2 | Why we do not define sample completeness as the proportion of detected diversity

In our framework (Equations 1 and 2), we have formulated sample completeness of order q as the proportion of the detected qth power sum. In the special case of $q = 0$, our formula reduces to the proportion of the detected diversity of order 0 (i.e., the proportion of detected species richness). One may wonder whether we can directly generalize this special case to define sample completeness of order q as the proportion of detected diversity of any order $q \geq 0$. Although this generalization is intuitively appealing, the proportion of detected diversity for $q = 2$ can be greater than unity. We give a simple example as follows. Consider a simple two-species case: assume the relative abundances for the two species are 0.1 and 0.9. The true diversity of order 2 is $1/(0.01 + 0.81) = 1.2195$. However, a million simulation trials for a sample size of 10 yield an average of 1.2295 for the detected diversity of order 2. Therefore, the proportion of detected diversity of order 2, for this specific example, can take a value greater than unity. A similar conclusion is also valid for other orders. For example, the true diversity of orders $q = 3$ and $q = 4$ are, respectively, 1.1704 and 1.1508, whereas the corresponding detected diversity values in a sample of size 10 based on a million simulations are 1.1893 and 1.1702. This example shows that the intuitive generalization is not a proper way to define sample completeness.

7.3 | Future research

In this paper, comparisons of diversity based on asymptotic and non-asymptotic approaches have been performed for each assemblage separately, followed by comparisons of within-assemblage curves across multiple assemblages. Beta diversity refers to the extent of differentiation or dissimilarity in species composition among a set of assemblages in a geographical area, over a time period, or along an environmental gradient. There are many concepts and measures of beta diversity and related similarity indices. The variance framework (derived from the total variance of an assemblage species abundance matrix) and diversity decomposition (based on partitioning gamma diversity into alpha and beta components) are two major approaches. Chao and Chiu (2016) bridged the two approaches and showed that they converge to the same classes of (dis)similarity measures. By analogy to within-assemblage diversity, beta diversity and (dis)similarity measures also depend on sample size and sample completeness. Thus, a worthwhile research topic would be to develop asymptotic and non-asymptotic methodologies to make fair comparisons for (dis)similarity measures across multiple sets of assemblages.

8 | CONCLUSION

In this paper, we have proposed a class of measures parameterized by an order $q \geq 0$ to quantify the sample completeness of a biological survey, based on species abundance or incidence data. All the theoretical formulas and their estimators for a general order q and three special orders ($q = 0$, 1 and 2) are summarized in Table 1. To compare diversity and evenness across assemblages based on incomplete data, we have proposed a four-step analysis procedure in Section 5. Our four-step procedure links sample completeness, diversity estimation, rarefaction and extrapolation, and evenness in a fully integrated approach. We have applied the four-step analysis procedure to four contrasting field data examples. We recommend future studies use the four-step analysis to assess sample completeness and compare diversity and evenness. The updated online software iNEXT-4steps is available to facilitate all computations.

ACKNOWLEDGEMENTS

This paper is a contribution of the symposium “Biodiversity estimation in time and space: a macroecological perspective of community assembly and diversity patterns” organized by Yasuhiro Kubota and held on October 24, 2018 at the University of the Ryukyus, Okinawa, Japan. Funding for the symposium and a related discussion group was provided by the University of the Ryukyus. This study was also supported by the Program for Advancing Strategic International Networks to Accelerate the Circulation of Talented Researchers and JSPS KAKENHI (no. 16H02755). We thank two reviewers for helpful and thoughtful comments and suggestions. We also thank the individuals and institutions supporting regional and national nodes of the Ocean Biogeographic Information System and Global Biodiversity Information Facility for making data freely available online without which the present research would not have been possible. This work was supported by the

Taiwan Ministry of Science and Technology under Contract 107-2118-M007-006-MY3 (to Anne Chao); and grants from the Research Grants Council of the Hong Kong Special Administrative Region, China (HKU 709413P, HKU 17303115, HKU 17302518) and the Seed Funding Programme for Basic Research of the University of Hong Kong (201111159140, 201611159053, 201711159057) (to Moriaki Yasuhara).

CONFLICT OF INTEREST

The authors declare no conflict of interest.

ORCID

Anne Chao https://orcid.org/0000-0002-4364-8101
Yasuhiro Kubota https://orcid.org/0000-0002-5723-4962
David Zelený https://orcid.org/0000-0001-5157-044X
Chun-Huo Chiu https://orcid.org/0000-0002-7096-2278
Ching-Feng Li https://orcid.org/0000-0003-0744-490X
Buntarou Kusumoto https://orcid.org/0000-0002-5091-3575
Moriaki Yasuhara https://orcid.org/0000-0003-0990-1764
Simon Thorn https://orcid.org/0000-0002-3062-3060
Chih-Lin Wei https://orcid.org/0000-0001-9430-0060
Mark J. Costello https://orcid.org/0000-0003-2362-0328
Robert K. Colwell https://orcid.org/0000-0002-1384-0354

REFERENCES

Asaad, I., Lundquist, C. J., Erdmann, M. V., & Costello, M. J. (2018). Delineating priority areas for marine biodiversity conservation in the coral triangle. *Biological Conservation*, *222*, 198–211. https://doi.org/10.1016/j.biocon.2018.03.037

Briggs, J. C. (2003). Marine centres of origin as evolutionary engines. *Journal of Biogeography*, *30*, 1–18. https://doi.org/10.1046/j.1365-2699.2003.00810.x

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

Chao, A. (1987). Estimating the population size for capture-recapture data with unequal catchability. *Biometrics*, *43*, 783–791. https://doi.org/10.2307/2531532

Chao, A., & Chiu, C.-H. (2016). Bridging two major approaches (the variance framework and diversity decomposition) to beta diversity and related similarity and differentiation measures. *Methods in Ecology and Evolution*, *7*, 919–928. https://doi.org/10.1111/2041-210x.12551

Chao, A., Chiu, C.-H., Colwell, R. K., Magnago, L. F. S., Chazdon, R. L., & Gotelli, N. J. (2017). Deciphering the enigma of undetected species, phylogenetic, and functional diversity based on Good-Turing theory. *Ecology*, *98*, 2914–2929. https://doi.org/10.1002/ecy.2000

Chao, A., & Colwell, R. K. (2017). Thirty years of progeny from Chao's inequality: Estimating and comparing richness with incidence data and incomplete sampling (invited article). *Statistics and Operation Research Transactions*, *41*, 3–54. https://doi.org/10.2436/20.8080.02.33

Chao, A., Gotelli, N. G., Hsieh, T. C., Sander, E. L., Ma, K. H., Colwell, R. K., & Ellison, A. M. (2014). Rarefaction and extrapolation with Hill numbers: A framework for sampling and estimation in species biodiversity studies. *Ecological Monographs*, *84*, 45–67. https://doi.org/10.1890/13-0133.1

Chao, A., & Jost, L. (2012). Coverage-based rarefaction and extrapolation: Standardizing samples by completeness rather than size. *Ecology*, *93*, 2533–2547. https://doi.org/10.1890/11-1952.1

Chao, A., & Jost, L. (2015). Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution*, *6*, 873–882. https://doi.org/10.1111/2041-210x.12349

Chao, A., & Ricotta, C. (2019). Quantifying evenness and linking it to diversity, beta diversity, and similarity. *Ecology*, *100*, e02852. https://doi.org/10.1002/ecy.2852

Chiou, C.-R., Hsieh, C.-F., Wang, J.-C., Chen, M.-Y., Liu, H.-Y., Yeh, C.-L., … Song, M. G.-Z. (2009). The first national vegetation inventory in Taiwan. *Taiwan Journal of Forest Science*, *24*, 295–302.

Colwell, R. K., & Chao, A. (2020). Measuring and comparing class diversity in archaeological assemblages: A brief guide to the history and state-of-the-art in diversity statistics. In M. I. Eren & B. Buchanan (Eds.), *Defining and measuring diversity in archaeology*. New York: Berghahn.

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S. Y., Mao, C. X., Chazdon, R. L., & Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation, and comparison of assemblage. *Journal of Plant Ecology*, *5*, 3–21. https://doi.org/10.1093/jpe/rtr044

Colwell, R. K., & Coddington, J. A. (1994). Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London B—Biological Sciences*, *345*, 101–118. https://doi.org/10.1098/rstb.1994.0091

Colwell, R. K., Mao, C. X., & Chang, J. (2004). Interpolating, extrapolating, and comparing incidence-based species accumulation curves. *Ecology*, *85*, 2717–2727. https://doi.org/10.1890/03-0557

Costello, M. J., Tsai, P., Wong, P. S., Cheung, A., Basher, Z., & Chaudhary, C. (2017). Marine biogeographic realms and species endemicity. *Nature Communications*, *8*(1057), 1–10. https://doi.org/10.1038/s41467-017-01121-2

Ellison, A. M. (2010). Partitioning diversity. *Ecology*, *91*, 1962–1963. https://doi.org/10.1890/09-1692.1

Good, I. J. (1953). The population frequencies of species and the estimation of population parameters. *Biometrika*, *40*, 237–264. https://doi.org/10.1093/biomet/40.3-4.237

Good, I. J. (1983). *Good thinking: The foundations of probability and its applications*. Minneapolis: University of Minnesota Press.

Good, I. J. (2000). Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval enigma. *Journal of Statistical Computation and Simulation*, *66*, 101–111. https://doi.org/10.1080/00949650008812016

Good, I. J., & Toulmin, G. (1956). The number of new species and the increase of population coverage when a sample is increased. *Biometrika*, *43*, 45–63. https://doi.org/10.1093/biomet/43.1-2.45

Gotelli, N. J., & Chao, A. (2013). Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In S. Levin (Ed.), *Encyclopedia of biodiversity* (2nd ed., pp. 195–211). Waltham MA: Academic Press.

Gotelli, N. J., & Colwell, R. K. (2001). Quantifying biodiversity: Procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters*, *4*, 379–391. https://doi.org/10.1046/j.1461-0248.2001.00230.x

Gotelli, N. J., & Colwell, R. K. (2011). Estimating species richness. In A. Magurran & B. McGill (Eds.), *Biological diversity: Frontiers in measurement and assessment* (pp. 39–54). Oxford: Oxford University Press.

Hill, M. O. (1973). Diversity and evenness: A unifying notation and its consequences. *Ecology*, *54*, 427–432. https://doi.org/10.2307/1934352

Hsieh, T. C., Ma, K. H., & Chao, A. (2016). iNEXT: An R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution*, *7*, 1451–1456. https://doi.org/10.1111/2041-210x.12613

Jost, L. (2010). The relation between evenness and diversity. *Diversity*, *2*, 207–232. https://doi.org/10.3390/d2020207

Kusumoto, B., Costello, M. J., Kubota, Y., Shiono, T., Wei, C.-L., Yasuhara, M., & Chao, A. (2020). *A quest for coral biodiversity shortfalls: Global-scale diversity estimation of tropical Scleractinia*. *Ecological Research, 35*, 315−326.

Li, C.-F., Chytrý, M., Zelený, D., Chen, M. -Y., Chen, T.-Y., Chiou, C.-R., … Hsieh, C.-F. (2013). Classification of Taiwan forest vegetation. *Applied Vegetation Science*, *16*, 698–719. https://doi.org/10.1111/avsc.12025

Magurran, A. E., & McGill, B. J. (Eds.). (2011). *Biological diversity: Frontiers in measurement and assessment*. Oxford: Oxford University Press.

McGrayne, S. B. (2011). *The theory that would not die: How Bayes' rule cracked the enigma code, hunted down Russian submarines, and emerged triumphant from two centuries of controversy*. New Haven: Yale University Press.

Pielou, E. C. (1966). The measurement of diversity in different types of biological collections. *Journal of Theoretical Biology*, *13*, 131–144. https://doi.org/10.1016/0022-5193(66)90013-0

Shin, C. P., Yasuhara, M., Iwatani, H., Kase, T., Hayashi, A. G. S. H., Kurihara, Y., & Pandita, H. (2019). Neogene marine ostracod diversity and faunal composition in Java, Indonesia: Indo-Australian Archipelago biodiversity hotspot and the Pliocene diversity jump. *Journal of Crustacean Biology*, *39*, 244−252. Doi:https://doi.org/10.1093/jcbiol/ruy110.

Su, H. -J. (1985). Studies on the climate and vegetation types of the natural forests in Taiwan (III) A Scheme of Geographical Climatic Regions. *Quarterly Journal of Chinese Forestry*, *18*, 33–44.

Thorn, S., Bässler, C., Svoboda, M., & Müller, J. (2017). Effects of natural disturbances and salvage logging on biodiversity - lessons from the bohemian Forest. *Forest Ecology and Management*, *388*, 113–119. https://doi.org/10.1016/j.foreco.2016.06.006

Thorn, S., Bußler, H., Fritze, M. -A., Goeder, P., Müller, J., Weiß, I., & Seibold, S. (2016). Canopy closure determines arthropod assemblages in microhabitats created by windstorms and salvage logging. *Forest Ecology and Management*, *381*, 188–195. https://doi.org/10.1016/j.foreco.2016.09.029

Veron, J. E. N., Devantier, L. M., Turak, E., Green, A. L., Kininmonth, S., Stafford-Smith, M., & Peterson, N. (2009). Delineating the coral triangle Galaxea. *Journal of Coral Reef Studies*, *11*, 91–100. https://doi.org/10.3755/galaxea.11.91

Veron, J. E. N., & Stafford-Smith, M. (2000). *Corals of the world* (Vol. 1–3). Townsville, Australia: Australian Institute of Marine Sciences.

SUPPORTING INFORMATION

Additional supporting information may be found online in the Supporting Information section at the end of this article.

How to cite this article: Chao A, Kubota Y, Zelený D, et al. Quantifying sample completeness and comparing diversities among assemblages. *Ecological Research*. 2020;35:292–314. https://doi.org/10.1111/1440-1703.12102