How should we estimate diversity in the fossil record? Testing richness estimators using sampling-standardised discovery curves

Close, Roger; Evers, Serjoscha; Alroy, John; Butler, Richard

DOI:
10.1111/2041-210X.12987

Document Version
Publisher's PDF, also known as Version of record

Citation for published version (Harvard):
Close, R, Evers, S, Alroy, J & Butler, R 2018, 'How should we estimate diversity in the fossil record? Testing richness estimators using sampling-standardised discovery curves', Methods in Ecology and Evolution, vol. 9, no. 6, pp. 1386-1400. https://doi.org/10.1111/2041-210X.12987

Received: 12 January 2018 | Accepted: 9 February 2018

DOI: 10.1111/2041-210X.12987

RESEARCH ARTICLE

Methods in Ecology and Evolution BRITISH ECOLOGICAL SOCIETY

How should we estimate diversity in the fossil record? Testing richness estimators using sampling-standardised discovery curves

Roger A. Close[1] | Serjoscha W. Evers[2] | John Alroy[3] | Richard J. Butler[1]

[1]School of Geography, Earth and Environmental Sciences, University of Birmingham, Edgbaston, Birmingham, UK

[2]Department of Earth Sciences, University of Oxford, Oxford, UK

[3]Department of Biological Sciences, Faculty of Science and Engineering, Macquarie University, Sydney, NSW, Australia

Correspondence
Roger A. Close
Email: r.a.close@bham.ac.uk

Handling Editor: Natalie Cooper

Abstract

1. To infer genuine patterns of biodiversity change in the fossil record, we must be able to accurately estimate relative differences in numbers of taxa (richness) despite considerable variation in sampling between time intervals. Popular subsampling (=interpolation) methods aim to standardise diversity samples by rarefying the data to equal sample size or equal sample completeness (=coverage). Standardising by sample size is misleading because it compresses richness ratios, thereby flattening diversity curves. However, standardising by coverage reconstructs relative richness ratios with high accuracy. Asymptotic richness extrapolators are widely used in ecology, but rarely applied to fossil data. However, a recently developed parametric extrapolation method, TRiPS (True Richness estimation using Poisson Sampling), specifically aims to estimate the true richness of fossil assemblages.
2. Here, we examine the suitability of a range of richness estimators (both interpolators and extrapolators) for fossil datasets, using simulations and a novel method for comparing the performance of richness estimators with empirical data. We constructed sampling-standardised discovery curves (SSDCs) for two datasets, each spanning 150 years of palaeontological research: Mesozoic dinosaurs at global scale, and Mesozoic–early Cenozoic tetrapods from North America. These approaches reveal how each richness estimator responds to both simulated best-case and empirical real-world accumulation of fossil occurrences.
3. We find that extrapolators can only truly standardise diversity data once sampling is sufficient for richness estimates to have asymptoted. Below this point, directly comparing extrapolated estimates derived from samples of different sizes may not accurately reconstruct relative richness ratios. When abundance distributions are not perfectly flat and sampling is moderate to good, but not perfect, TRiPS does not extrapolate, because it overestimates binomial sampling probabilities. Coverage-based interpolators, by contrast, generally yield more stable subsampled diversity estimates, even in the face of dramatic increases in face-value counts of species richness. Richness estimators that standardise by coverage are

among the best currently available methods for reconstructing deep-time biodiversity patterns. However, we recommend the use of sampling-standardised discovery curves to understand how biased reporting of fossil occurrences may affect sampling-standardised diversity estimates.

KEYWORDS
Dinosauria, diversity, extrapolation, fossil record, interpolation, sample coverage, shareholder quorum subsampling, species accumulation curve

1 | INTRODUCTION

Early studies of taxonomic richness through deep time (e.g. Benton, 1985; Sepkoski, Bambach, Raup, & Valentine, 1981; Valentine, 1969) interpreted the fossil record literally using face-value (=raw or observed) counts of taxa. However, because fossil record sampling varies considerably among clades, geological time-intervals and geographic regions, direct comparisons of face-value richness can be misleading (e.g. Alroy et al., 2001, 2010a, 2010b; Peters, 2005; Raup, 1972; Smith & McGowan, 2011). To infer genuine patterns of deep-time biodiversity, we need methods that successfully standardise samples of unequal sizes and permit direct comparisons of richness among assemblages.

An early approach was to standardise samples by size, drawing down samples to equal numbers of specimens, individuals or localities (classical rarefaction [CR]; Sanders, 1968). However, item-quota standardisation methods such as CR under-sample more diverse assemblages, compressing relative richness ratios and artificially flattening diversity curves (Alroy, 2010b,c; Chao & Jost, 2012). The solution to this problem is to standardise samples to equal levels of completeness, or "coverage" of the species' underlying frequency distribution (Alroy, 2010a; Chao & Jost, 2012; Jost, 2010). This approach is known among palaeobiologists as shareholder quorum subsampling (SQS), and among ecologists as coverage-based rarefaction (CBR). It reconstructs richness ratios with high accuracy, provided that the shape of the abundance distribution does not vary substantially between assemblages (Alroy, 2010a–2010c; Chao & Jost, 2012).

Asymptotic richness extrapolators use relative frequencies of rare species to analytically estimate undetected species from limited samples (e.g. Chao1/2, Chao, 1984, 1987; ACE, Chao, 2005; and jackknife, Burnham & Overton, 1978). Extrapolators are widely used in ecology (Chao & Chiu, 2016; Gotelli & Chao, 2013), but only rarely for fossil data (e.g. Vavrek & Larsson, 2010). However, a parametric extrapolator, TRiPS (True Richness estimation using Poisson Sampling), was recently proposed for fossil data (Starrfelt & Liow, 2016a).

Here, we describe a new approach for examining the real-world performance of richness estimators when confronted with new data. We evaluate the ability of both interpolators and extrapolators to successfully standardise diversity data and accurately reconstruct relative magnitudes of richness between assemblages. We construct sampling-standardised discovery curves (SSDCs; also known as species-accumulation or collector curves) spanning 150 years of palaeontological exploration for (1) Mesozoic–early Cenozoic terrestrial tetrapods, and (2) Mesozoic dinosaurs. This novel historical dimension reveals how the potentially biased accumulation of new fossil data affects richness estimates generated by different methods. We interpret empirical patterns in light of results from simulated datasets, in which richness and evenness are precisely and independently varied and sampling is unbiased. Although we focus on fossil datasets, many of our conclusions are equally applicable to modern-day ecological studies.

2 | MATERIALS AND METHODS

2.1 | Richness estimators

2.1.1 | Interpolators

We evaluated two interpolation methods, CR and SQS. CR is flawed because it artificially compresses richness ratios (Alroy, 2010b, 2010c; Chao & Jost, 2012), and we implemented it in our simulations for illustrative reasons only (using the R package iNEXT; Hsieh, Ma, & Chao, 2016). The alternative approach of standardising data to equal coverage (="quorum" level) was proposed and implemented algorithmically by Alroy (2009, 2010a, 2010b, 2010c) under the name SQS, and independently described by Jost (2010). Chao and Jost (2012) described the analytical solutions and expanded the approach to permit extrapolation, calling it coverage-based rarefaction (CBR). The names SQS and CBR refer to the same broad approach of standardising diversity samples by coverage, and do not uniquely map onto any method of implementation (algorithmic or analytical) or piece of software (e.g. J. ALROY'S SQS Perl and R scripts, and iNEXT [Hsieh et al., 2016]; see Appendix S1 for detailed discussion of SQS).

Coverage is an objective measure of diversity-sample completeness that can be efficiently estimated from the frequencies of rare species in a sufficiently large sample (Esty, 1986; see Appendix S4). The simplest and most commonly used coverage estimator is Good's *u* (Good, 1953). Good's *u* ranges between 0 and 1, and is equal to one minus the number of singletons (species only observed once) divided by the total number of sampling units (specimens, individuals or occurrences). Sampling is poor when there are many singletons, and good when there are few. Coverage indicates what percentage of individuals in the original population belong to species included in the sample. Conversely, the complement of coverage, the 'coverage deficit', indicates the fraction of individuals in the source population belonging to unsampled species. The coverage deficit at any particular level of sample completeness is proportional to the slope at that point on a

rarefaction curve; this also corresponds to the probability that a new species will be observed by adding one more individual to the sample (Chao & Jost, 2012). For example, if coverage is estimated to be 0.9, the species in the sample account for 90% of the individuals in the focal assemblage, and there is a 10% chance that a new species will be discovered if the sample size is increased by one (Chao & Jost, 2012).

Shareholder quorum subsampling has been implemented using two subsampling algorithms (Alroy, 2009, 2010a, 2010b, 2010c, 2014) and one set of analytical equations (Chao & Jost, 2012; see Appendix S1). Unlike the original approximate algorithm, the exact algorithm (Alroy, 2014) used here consistently imposes the target quorum. During each subsampling trial, all occurrences are drawn sequentially and randomly, continually tracking the value of Good's *u*. As occurrences are drawn, *u* may either rise or fall. Each time *u* crosses the target quorum, richness is recorded, and the median of these values from all subsampling trials represents the overall estimate. The exact algorithm produces results that are identical to the analytical equations of Chao and Jost (2012, implemented in iNEXT; see Figure S1). Importantly, the exact algorithm lets us implement additional protocols to address biases affecting fossil occurrence datasets (Alroy, 2009, 2010a, 2010b, 2010c, 2014; see Appendix S1).

For the simulations, we used a combination of the analytical equations and the exact algorithm, newly implemented here in the R language. SQS richness estimates for fossil datasets were calculated, using iNEXT and the exact algorithm implemented in JA's Perl script version 4.3. The latter allows us to apply the three-collections-per-reference protocol necessary to account for the reference effect (see Appendix S1).

2.1.2 | Extrapolators

We evaluated three extrapolators: TRiPS (Starrfelt & Liow, 2016a), Chao1 (Chao, 1984) and λ^5 ("lambda-5"; Alroy, 2017). TRiPS aims to estimate true richness by modelling per-interval sampling rates for extinct lineages as a homogeneous Poisson process. Maximum likelihood is used to infer a single sampling rate for all taxa present in each interval, using observed taxon occurrence frequencies and interval durations. Sampling rates are then used to estimate a single per-lineage binomial probability per interval: the odds that a species would be sampled given that it was extant during that interval. The richness estimated by TRiPS is that which maximises the binomial likelihood given that binomial sampling probability and the observed number of species. We implemented TRiPS using the R scripts provided by Starrfelt and Liow (2016a, 2016b).

Chao1 is an asymptotic richness extrapolator widely used in ecology (Colwell & Coddington, 1994; Gotelli & Chao, 2013) that uses information about rare species (singletons and doubletons) to estimate a lower bound for true species richness. Chao1 assumes that singletons, doubletons and undetected species have equal underlying frequencies, and that the sample size is large enough that the mean abundances of singletons and undetected species are similar (Chao & Chiu, 2016). Estimates should therefore be more downward-biased as communities become more uneven and sample sizes are smaller.

To address this, Alroy (2017) proposed λ^5, which reformulates Chao1 in terms of Poisson sampling, and incorporates information about the number of observed species, singletons and total individuals sampled. The λ^5 equation is solved via a simple hill-climbing algorithm. Although λ^5 also assumes a flat abundance distribution, it incorporates information about sample size to reduce the downward bias when abundance distributions are uneven.

2.2 | Simulation experiments

We performed three simulation experiments to show how richness estimators perform under ideal conditions, when true richness and evenness are known and sampling is unbiased. The first two experiments (SE1 and SE2) precisely and independently varied sampling effort, true richness and evenness, while the third (SE3) varied sampling effort systematically, but varied true richness and evenness stochastically.

For SE1 and SE2, we simulated communities with all combinations of four values of true richness (50, 100, 200 and 400 species) and four levels of underlying evenness (one perfectly even/flat, and three lognormally distributed, with standard deviations [*SDs*] of 1, 1.5 and 2). The simulations in SE1 are directly analogous to sampling-standardised discovery curves derived from empirical datasets. For each simulated community, samples were drawn progressively at sample sizes ranging from 1 to 10,000 individuals. At each sample size, we recorded face-value species counts and richness estimates from SQS, CR, TRiPS, Chao1 and λ^5. This procedure produces a single simulated discovery curve (face-value species counts) and set of SSDCs (extrapolated richness estimates). The procedure was repeated 1,000 times and the curves averaged to produce rarefaction curves for each richness estimator. For face-value species counts, this procedure yields an item-quota or size-based rarefaction curve (i.e. a CR curve). However, performing this procedure for other richness estimators yields sampling-standardised rarefaction curves (SSRCs), and allows point estimates using size-rarefied extrapolated richness estimates (e.g. size-rarefied Chao1, TRiPS or λ^5 richness). Such curves reveal how each richness estimator is expected to respond to the progressive accumulation of data when sampling is entirely random and unbiased.

The simulations in SE2 also generated rarefaction curves for the simulated communities. However, these simulations are not analogous to empirical SSDCs, because the data were rarefied by coverage, not sample size. SE2 therefore demonstrates how each richness estimator is expected to respond to progressive increases in sample coverage (from 0.1 to 0.9999). Although the analytical equations of Chao and Jost (2012) permit Chao1 estimates at specific levels of coverage, we opted to use the exact algorithm because this allows us to standardise any estimator to equal coverage. We achieved this by modifying the code for the exact algorithm in R to calculate not only simple counts of species whenever the target quorum was crossed or reached, but also estimates from Chao1 and λ^5, yielding coverage-rarefied extrapolated richness estimates. The asymptotic richness

estimates from these extrapolators are derived from repeated subsamples of the data at specific levels of coverage.

We did not rarefy TRiPS to equal coverage because implementing this method within the exact algorithm was too computationally intensive. TRiPS runs approximately three orders of magnitude slower than other extrapolators, and the exact algorithm used to standardise richness estimators to equal coverage is also computationally demanding. However, results from size-based rarefaction of TRiPS richness estimates in SE1 demonstrate that coverage-based rarefaction of TRiPS would not be beneficial (see Section 3).

SE3 tested the sensitivity of richness estimators to stochastic variation in richness and evenness. We generated sampling-standardised richness estimates for many simulated communities in which true richness was sampled from a lognormal distribution (*SD* = 1 and *M* = 5 on a log scale), and the *SD* of the underlying lognormal abundance distribution was randomly varied from 1 to 2 on a log scale. These were standardised at both a range of sample sizes and coverages.

We also used the simulation framework from SE1 to show expected counts of singletons, doubletons and multitons (species that have been sampled at least twice) with increasing sampling effort. Comparing curves of singletons, doubletons and multitons from empirical fossil datasets to expected patterns under entirely unbiased sampling can shed light on the nature of reporting biases. For example, novelty biases (see Section 4) are expected to inflate the frequencies of singletons relative to multitons, and might therefore distort curves of counts as a function of sampling effort.

2.3 | Empirical sampling-standardised discovery curves

Full details of the fossil occurrence data are provided in Appendix S2. We downloaded Mesozoic–early Eocene occurrence data for Tetrapodomorpha, and Mesozoic occurrence data for Dinosauromorpha, from the Paleobiology Database (PaleoDB). Marine tetrapods and flying taxa were excluded, and the datasets were cleaned (see Appendix S2).

Our analyses focus on two partitions of these data. The first comprises North American data because this continent has the best sampled fossil record for much of this interval. The second partition comprises all global occurrences of Mesozoic dinosaurs. These data are exceptionally complete and well-vetted in the PaleoDB, and there has been intense interest in reconstructing dinosaur diversity patterns (e.g. Barrett, McGowan, & Page, 2009; Butler, Benson, Carrano, Mannion, & Upchurch, 2011; Starrfelt & Liow, 2016a; Tennant, Chiarenza, & Baron, 2018; Upchurch, Mannion, Benson, Butler, & Carrano, 2011), including in the initial publication of TRiPS. Global diversity curves suffer from profound issues with highly-variable sampling universes (Appendix S2), and here we only analyse dinosaur data at global level to enable direct comparison with Starrfelt and Liow (2016a).

We reconstructed SSDCs for each geological stage-level time interval by subsetting our data to create 150 historical datasets, representing yearly timeslices through the history of palaeontological discovery, from 1866 to 2016. Each PaleoDB occurrence is associated with a published reference that corresponds to either the original description or the latest accepted taxonomic revision. This information was used to limit each historical timeslice to only those occurrences published prior to or during the year in question. Historical snapshots of the fossil record may include taxonomic opinions and identifications that were later rendered obsolete. This provides a more accurate picture of the history of palaeontological discovery, and is akin to using historical literature compilations (e.g. Alroy, 2000; Sepkoski, 1993; Tennant et al., 2018). We did not construct SSDCs using CR because our simulations provide further evidence that the method produces misleading results (see below).

Sampling-standardised discovery curves in which collection effort is quantified by time in years may be misleading if discovery rates are strongly heterogeneous. We therefore focus on SSDCs in which effort is quantified by the chronological addition of occurrences. Together with coverage estimates, these provide a much clearer view of sampling effort through collector-time.

To examine biases in the real-world accumulation of species in the fossil record, we compared empirical SSDCs to null distributions where the order in which occurrences are discovered is repeatedly randomised (these are equivalent to sampling-standardised rarefaction curves). This is achieved by generating many replicate datasets in which publication years for occurrences are randomly assigned a year and SSDCs are calculated. These null distributions shed light on the performance of sampling-standardisation methods for constructing SSDCs in the absence of systematic collection and reporting biases (Alroy, 2010a, 2010b, 2010c), including the expansion of the sampling universe (e.g. when the empirical SSDC falls well above or below the range observed in the null). We calculated palaeogeographic spread (the spatial distribution of fossil localities within a time interval) in order to quantify the expansion of the geographic sampling universe through collector-time (Appendix S2). All analyses were conducted in R (version 3.2.2; R Development Core Team, 2015), unless otherwise stated. All analysis code and data have been archived on Zenodo (https://doi.org/10.5281/zenodo.1167536; Close et al., 2018).

3 | RESULTS

3.1 | Simulation experiments

Rarefying face-value species counts by sample size, as in SE1, produces a "classical" rarefaction curve (i.e. one showing how CR estimates change with sample size), while rarefying face-value species counts by coverage, as in SE2, yields a coverage-based rarefaction curve (i.e. showing how SQS estimates change with coverage). When sampling is unbiased, interpolated SQS and CR estimates are—by their very nature—relatively insensitive to sheer sample-size: the data is either sufficient to provide an estimate at the desired standardisation level, or it is not (Figures 1 [i.e. CR], 2 and S2 [i.e. SQS]). As a result, SSRCs for SQS and CR are simply flat lines extending out

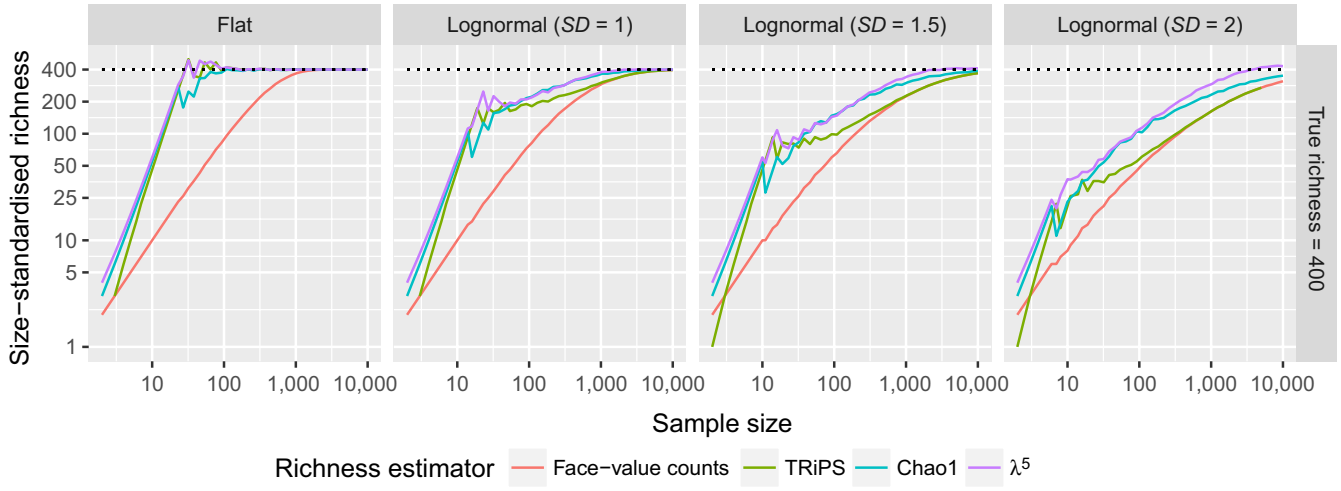


FIGURE 1 Size-based rarefaction curves for face-value counts (=CR), Chao1 and λ^5, analysing communities from Simulation Experiment 1. Columns represent true evenness values. Until they asymptote, extrapolated richness estimates are strongly sample-size dependent. Extrapolators converge on true richness rapidly when communities are perfectly even, but take progressively longer as evenness decreases. When communities are not perfectly even, TRiPS ceases to extrapolate above a certain sample size. λ^5 performs better than other methods tested when communities are uneven

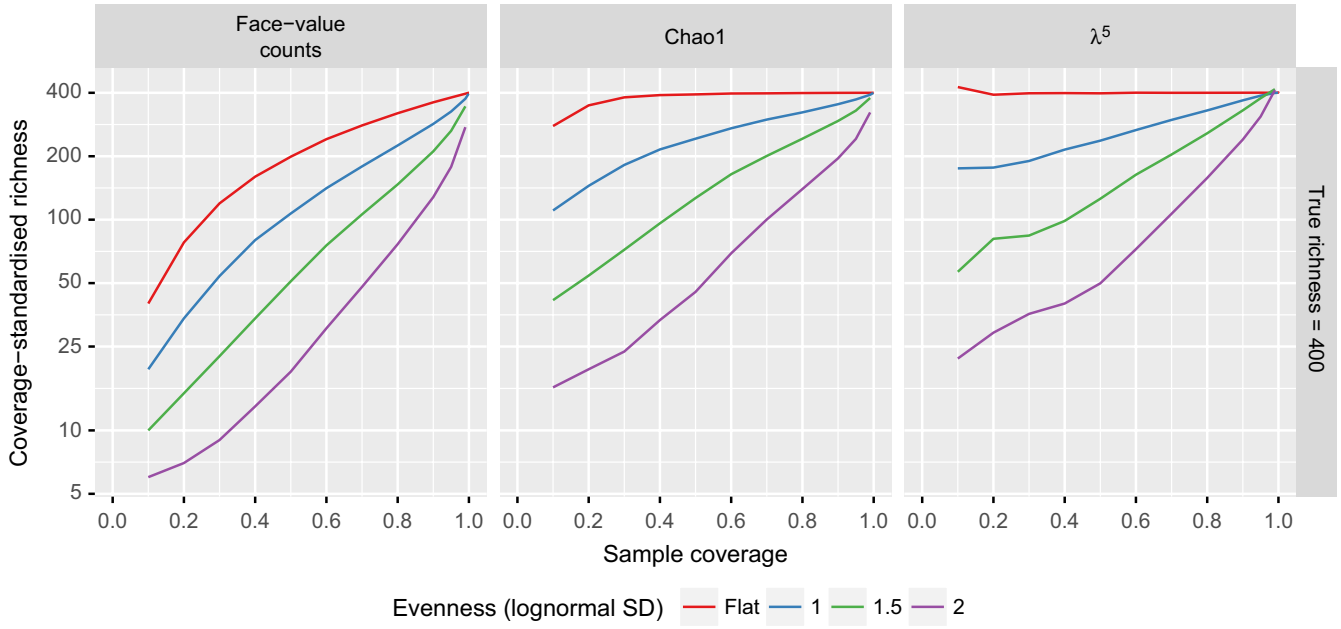


FIGURE 2 Coverage-based rarefaction curves using both face-value counts (=SQS) and extrapolated richness from Chao1 and λ^5, analysing communities from Simulation Experiment 1. TRiPS is not included due to computational issues, but Figure 1 demonstrates that this method performs poorly when abundance distributions are uneven and sampling is moderate to good

from the point at which the sample size or coverage is sufficient to obtain an estimate (Figure S3).

However, although it is relatively insensitive to raw sample size, CR artificially compresses richness ratios by progressively underestimating relative richnesses of more diverse communities (Alroy, 2010b, 2010c; Chao & Jost, 2012). This results in a nonlinear relationship between true and estimated richness (especially when evenness is high; Figure 3). In contrast, standardising by coverage yields perfectly accurate relative richnesses provided that the shape of the abundance distribution does not vary among communities (Figures S2, 2 and S3). As a result, coverage-standardised richness scales linearly with true richness (Figure 4).

By contrast, sampling-standardised rarefaction curves from SE1 (Figures 1, S3 and S4) show that below a threshold sampling level, extrapolated estimates from Chao1, λ^5 and TRiPS depend strongly on sample size. Richness estimates only asymptote on true richness and become insensitive to sample size once a threshold level of coverage has been met.

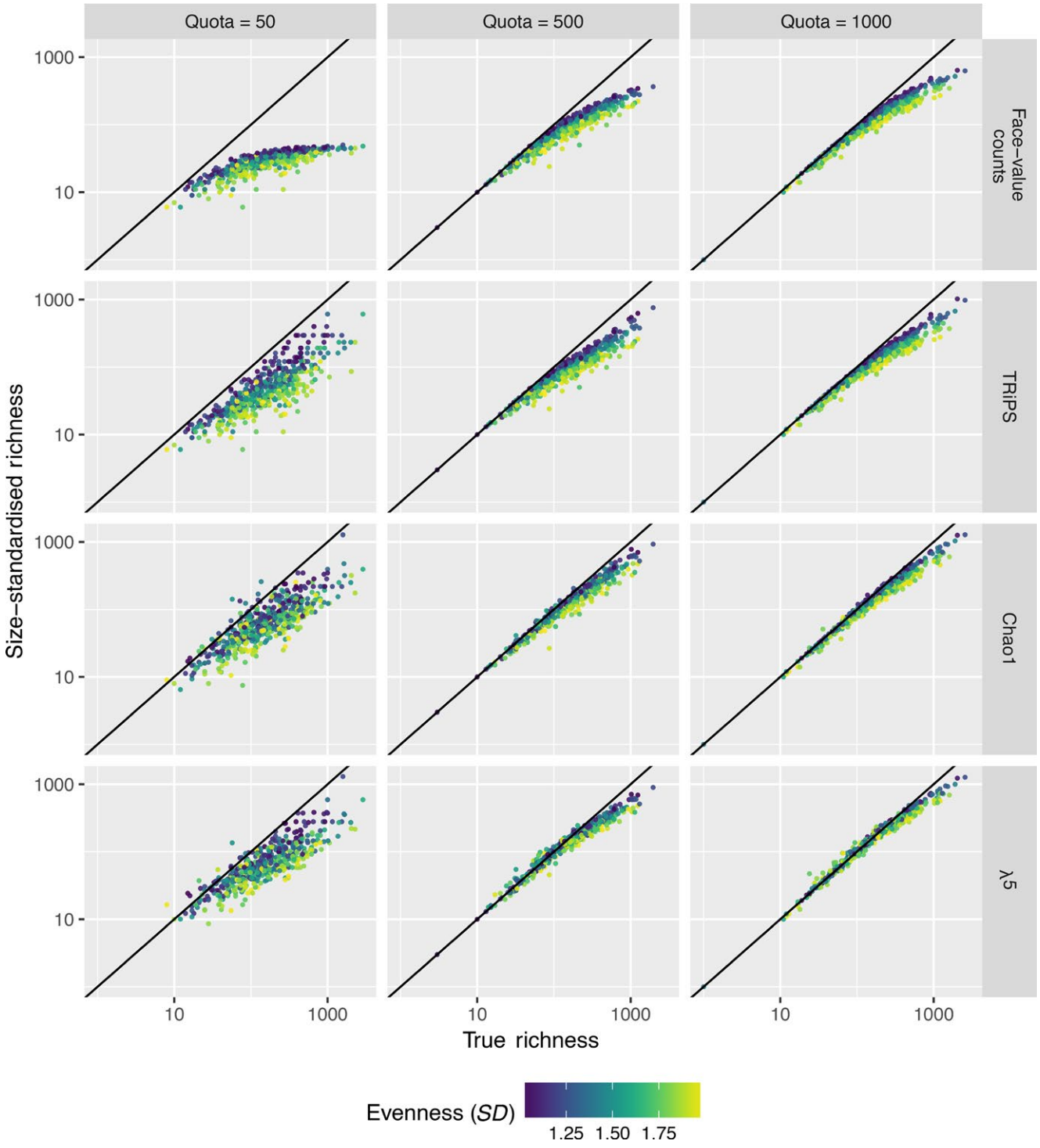


FIGURE 3 Relationship between true and estimated richness for estimators standardised by sample size (face-value counts [=CR], TRiPS, Chao1 and λ^5), analysing communities from Simulation Experiment 3. Standardising to equal sample size causes estimators to scale nonlinearly with true richness, particularly when sampling is limited. Standardising Chao1 and λ^5 to equal sample size yields a tighter relationship, but the nonlinear pattern remains

The sampling level required for extrapolators to asymptote becomes greater as true richness increases or evenness decreases. When communities are perfectly even—i.e. when the species abundance distribution is perfectly flat—sampling-standardised rarefaction curves for Chao1, TRiPS and λ^5 stabilise at very small sample sizes (Figure S3). For a perfectly even community with 400 species, these extrapolators asymptote on true richness after sample sizes reach 100 individuals, or when sample coverage is <0.5 (Figures S3 and 2). By contrast, face-value counts of species only asymptote on true richness after at least 1,000 individuals have been sampled, the point at

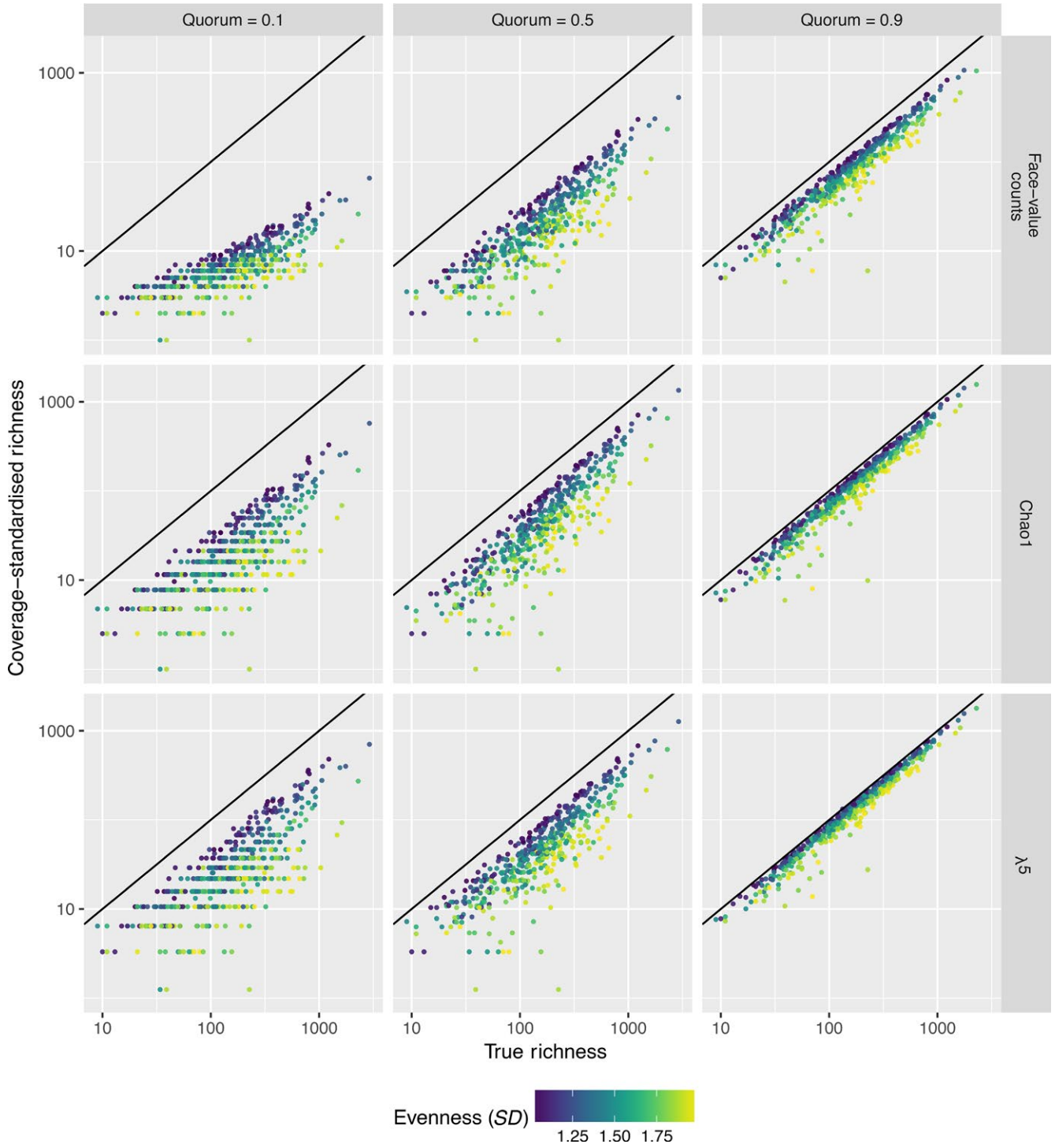


FIGURE 4 Relationship between true and estimated richness when standardising face-value counts (=SQS), Chao1 or λ^5 estimates to equal coverage, analysing communities from Simulation Experiment 3. Coverage-standardised estimators scale linearly with true richness. Variation in evenness (*SD* of underlying lognormal distribution) causes a looser relationship at lower levels of coverage, but the effect diminishes as coverage increases. Standardising extrapolators (especially λ^5) to equal coverage yields a visibly tighter relationship

which coverage is total. Confidence intervals are large at small sample sizes, with an upper bound that peaks sharply (greatly exceeding true richness) before shrinking, then disappearing as the coverage deficit diminishes to zero (Figure S4). Chao1 yields the most conservative estimate, but converges on true richness slightly later than TRiPS and λ^5.

As evenness decreases, extrapolators require progressively more data in order to converge on true richness (Figures S3–2). For communities with lognormal frequency distributions, λ^5 converges on true richness earlier than Chao1 (*SD* = 1–1.5), but initially overshoots true richness when evenness is very low (*SD* = 2). As evenness

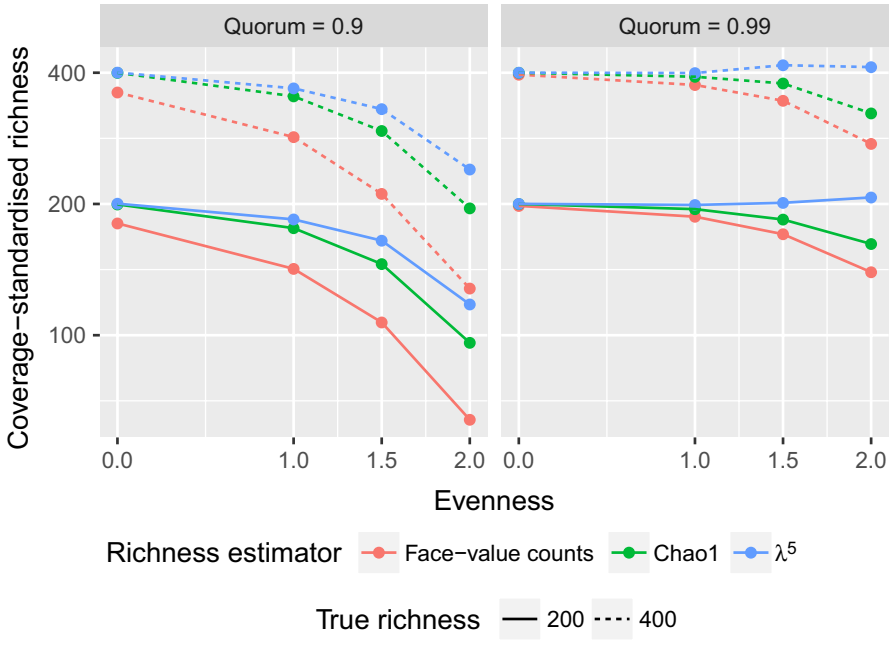


FIGURE 5 Effect of evenness on estimated richness when standardising face-value counts (=SQS), Chao1 or λ^5 estimates to equal coverage, analysing communities from Simulation Experiment 1. Estimators are superimposed to emphasise disparities in their response to evenness. Standardising extrapolators (especially λ^5) to equal coverage reduces the downward bias that results from low evenness

decreases, upper confidence interval bounds for Chao1 and λ^5 (but not TRiPS) usually approach or encompass true richness (Figure S4).

TRiPS, however, ceases to extrapolate (simply returning face-value counts of species) when the underlying frequency distribution is not perfectly flat and sample sizes are moderate to large (Figures 1 and S3). Once TRiPS ceases to extrapolate, confidence intervals shrink to negligible sizes (Figure S4). This even occurs when the underlying lognormal frequency distribution is comparatively even (*SD* = 1; Figure S3). When evenness is very low (*SD* = 2) and sample sizes are large, TRiPS ceases to yield richness estimates altogether.

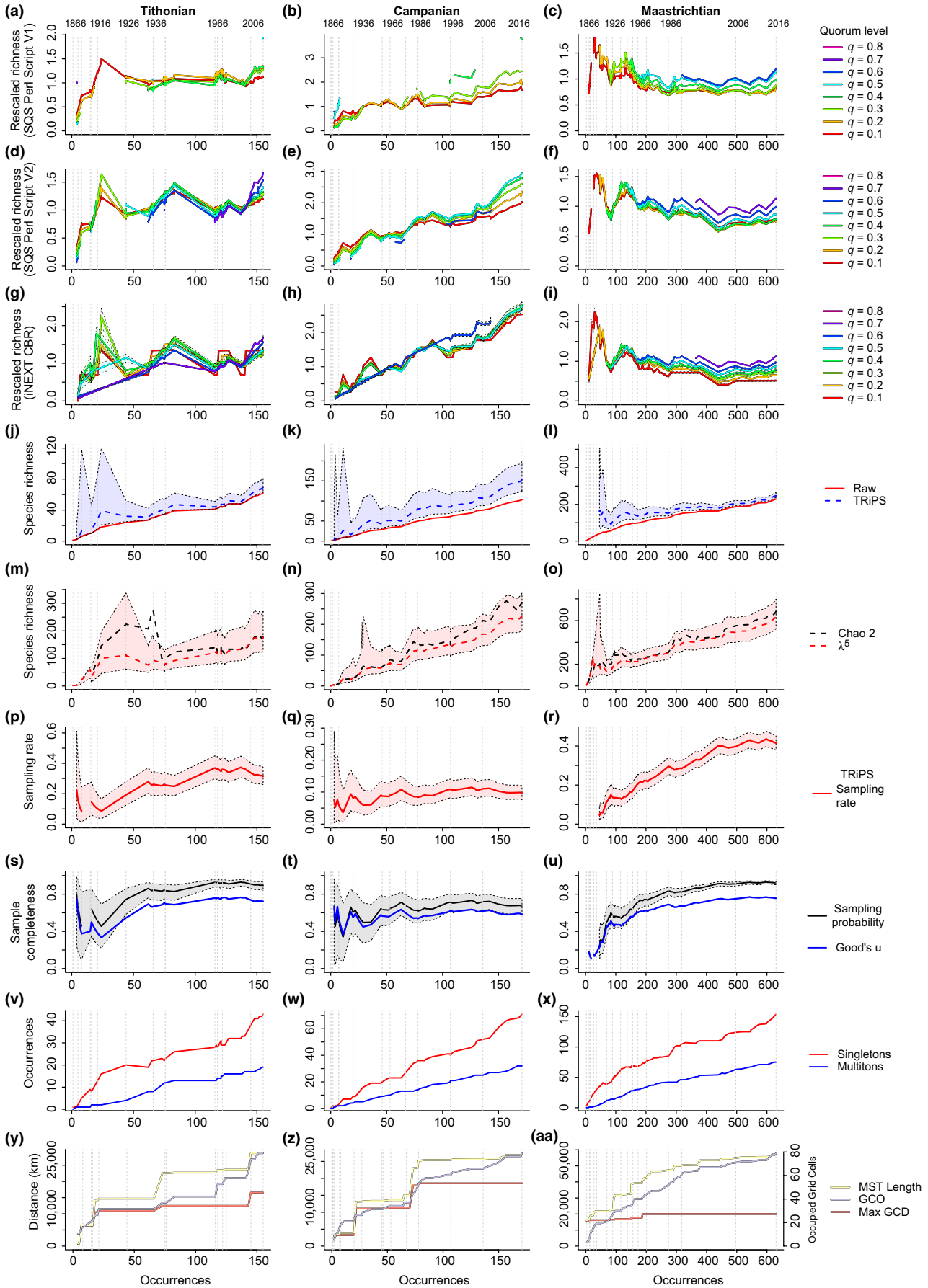
If sampling is limited relative to true richness, all richness estimators tested here are biased by low evenness. Regardless of whether samples are rarefied by size (CR) or coverage (SQS), richness estimates drop as evenness diminishes (compare Figures 5 and S5). When sampling is comparatively limited, substantial among-sample differences in evenness can severely confound estimates of relative richness: even at a quorum of 0.9, the coverage-standardised estimate for a community with 400 species and low evenness (*SD* = 2) is substantially less that for a perfectly even community with 200 species (Figure 5). As communities diverge in evenness, progressively greater coverage is required in order to accurately infer relative richness for communities as a whole, since it becomes ever harder to detect the rarest species. When sampling is very poor, standardising by coverage produces richness estimates that are slightly more biased by differences in evenness than standardising by sample size. However, when sampling is very good, the situation is reversed, and CR becomes more sensitive to evenness than SQS (Figure S6; see Discussion). The influence of evenness diminishes as sample size or coverage level increases (Figures 5, S5 and S7). Crucially, however, only standardising by coverage yields a linear relationship between true and estimated richness (Figures 3 and 4).

Downward biases to richness estimates caused by low evenness are substantially reduced by using coverage-based rarefaction of extrapolated richness estimates, rather than coverage-based rarefaction of simple face-value counts of species (=SQS/CBR). Coverage-rarefied λ^5 richness estimates are the least affected by evenness (compare richness estimates at a coverage of 0.99 for face-value counts, Chao1 and λ^5 in Figure 5; coverage-rarefied λ^5 is nearly unaffected by differences in evenness).

We were not able to coverage-rarefy TRiPS richness estimates because of computational issues (see Section 2). However, the results of simulations standardising TRiPS to equal sample size (Figures 1 and S3) show that the method only extrapolates when abundance distributions are flat or sampling is limited. Coverage-based rarefaction of TRiPS richness estimates would not alter this fact: with increasing sampling effort, coverage-rarefied TRiPS estimates for less-than-perfectly-even assemblages would simply converge on those from SQS.

Plotting counts of singletons, doubletons, tripletons and multitons obtained via simulation against sampling intensity (Figure S8)

FIGURE 6 Face-value (unstandardised) and sampling-standardised discovery curves for global dinosaur species during the Tithonian, Campanian and Maastrichtian. All variables visualised against chronologically added occurrences for reasons explained in Section 2. (a–c) SQS, using fossil-dataset protocols (see Appendix S1 for details). (d–f) SQS without fossil-dataset protocols. (g–i) SQS (iNEXT). (j-l) TRiPS (juxtaposed with face-value discovery curve). (m–o) Chao1 and λ^5. (p–r) TRiPS sampling rate. (s–u) TRiPS binomial sampling probability and Good's *u* (sampling rates are not expected to show a predictable relationship with sampling probabilities or coverage, but are shown in the same panels to allow easier comparison of patterns). (v–x) Counts of singleton and multiton species through collector-time (note that singletons generally accumulate faster than multitons due to reporting biases). (y-aa) Changes in palaeogeographic spread (summed minimum spanning tree length, occupancy of grid-cells of 2-degree latitude/longitude and maximum great circle distance)



Occurrences

reveals the patterns that should be expected if sampling is unbiased. Empirical patterns can be compared against these to assess the strength of the reporting biases (see below).

3.2 | Empirical sampling-standardised discovery curves

Empirical SSDCs using extrapolators show little sign of asymptoting (Figures S9j–o, 6j–o; Figures S10kk–hhh and S11y–nn). As our simulations predicted, TRiPS frequently tracks unstandardised discovery curves, particularly in well-sampled intervals (e.g. NAm tetrapods in the Maastrichtian, Danian and Ypresian, Figure S9j–l; and global dinosaurs during the Maastrichtian, Figure 6l). TRiPS stops extrapolating when the estimated per-lineage binomial sampling probability reaches 1, which should indicate that every lineage alive within the interval has been sampled. However, TRiPS often infers binomial sampling probabilities of 1 even when the coverage deficit is substantial, indicating that many species remain undetected. For example, in Ypresian tetrapods (Figure S9l), TRiPS sampling probability reached 1 in the early 1980s, yet face-value counts of species—and thus TRiPS richness—continued to climb.

Chao1 and λ^5 do consistently extrapolate, but estimates generally rise in step with new discoveries (Figures S9m–o and 6m–o). This suggests that fossil sampling is often insufficient for applying extrapolators to unstandardised data. One exception is in the Ypresian from 1996 to 2006, when numerous tetrapod occurrences containing few novel species were added (Figure S9o); here, only λ^5 increases. However, both Chao1 and λ^5 continue to rise from 2006 to 2016.

By contrast, subsampled richness does not consistently rise in step with new discoveries. We focus on SQS results using the three-collections-per-reference protocol and subsampled by collection ("V1"), but highlight how these protocols alter results. Provided a sufficient level of sampling has been achieved, SQS SSDCs are often remarkably stable despite substantial additions of data (Figures S9a–c, 6a–c). However, extrapolated estimates for North American tetrapods rise with new discoveries during the Maastrichtian, Danian and Ypresian, SQS richness changes little (Figure S9a–c). Similar patterns are evident in the global dinosaur dataset (Figures 6a–c, 7b–c and f–g). The stability of SQS SSDCs for Tithonian and Maastrichtian dinosaurs (Figure 6a,c) are especially worthy of note, as they are in stark contrast to the steep rises in unstandardised (face-value) and extrapolated curves.

Simulations show that SSDCs using subsampling methods will follow a perfectly flat trajectory if (1) sampling is random and unbiased, and (2) the size of the underlying sampling universe is static (Figure S3). However, idiosyncratic sampling of the fossil record may cause subsampled diversity estimates to fluctuate. Firstly, SQS richness may decline as new occurrences are added. This is most evident for global dinosaurs during the Maastrichtian (Figures 6c, 7b,f), where SQS richness for a quorum of 0.4 almost halves from 1976 to 2006 (when *c.* 1,000 occurrences were added). However, this decline disappears at a quorum of 0.5 (Figure 7c,g), and SQS richness stabilises at all quorum levels when 200–300 occurrences had accumulated (a level of sampling reached around 1980). Over the next three decades, the number of global Maastrichtian dinosaur occurrences doubled without affecting SQS richness. We attribute such declines to systematic biases in the reporting of fossil occurrences (see Section 4). Secondly, SQS SSDCs may rise after a period of stability (e.g. North American tetrapods at higher quorum levels during the Maastrichtian, Ypresian and Danian in the last two decades; Figure S9a–c). Coincident increases in the palaeogeographic spread of localities and in counts of singleton taxa (Figure S9v–x, y–aa) suggest that such rises are likely due to expansion of the sampling universe via exploration of previously unsampled regions (see Section 4).

Steep rises in SQS SSDCs are also common in the early phases of exploration (Figure 7b–c and f–g). For global dinosaurs at a quorum of 0.4, SQS SSDCs for most intervals rise steeply at first, only stabilising after over 200 occurrences have accrued (Figure 7b,f). This is most likely because coverage cannot be efficiently estimated with Good's *u* below this range of sample sizes. This is why a quality threshold of 20 references is commonly applied to filter unreliable SQS richness estimates (e.g. Benson et al., 2016), but SSDCs directly show when curves standardised estimates have stabilised.

The SQS Perl script with all fossil-dataset protocols disabled ("V2") produces nearly identical results to iNEXT (interpolated estimates only; Figures S9d–i and 6d–i). However, fossil-dataset protocols have a variable impact on SSDCs. Firstly, the three-collections-per-reference protocol often reduces the maximum obtainable quorum due to the concentration of occurrence data within monographic publications. Because the protocol limits the number of collections that may be drawn per subsampling trial to three per reference, it effectively caps the number of occurrences that can be drawn if some references contain many collections. This may lower the maximum attainable coverage. Secondly, the fossil-dataset protocols may alter SSDC patterns. In some intervals, these protocols have little effect (e.g. Maastrichtian dinosaurs; Figure 6c,f). However, curves for most intervals differ (e.g. Ypresian tetrapods, where the protocol eliminates a decline through 1996–2006 coincident with the addition of a monograph listing many new mammal occurrences, but few new taxa; Figure S9c,f).

Null distributions reveal the range of patterns SSDCs would take if all currently-known occurrences had been discovered in random order, and thus shed light on systematic reporting biases, such as a preference for reporting novel taxa, or systematic expansion of the sampling universe through collector-time (see Section 4). When the sampling universe is expanded late in collection history, the empirical SQS SSDC lies below the range of randomised collection histories (e.g. North American tetrapods during the Danian and Ypresian above quorum 0.1; Figure S12b,c). Progressively better sampling of the same universe causes the empirical curve to lie within the range of the null (e.g. for global dinosaurs during the Tithonian; Figure S13). Maastrichtian tetrapods exhibit both patterns depending on the quorum level (Figure

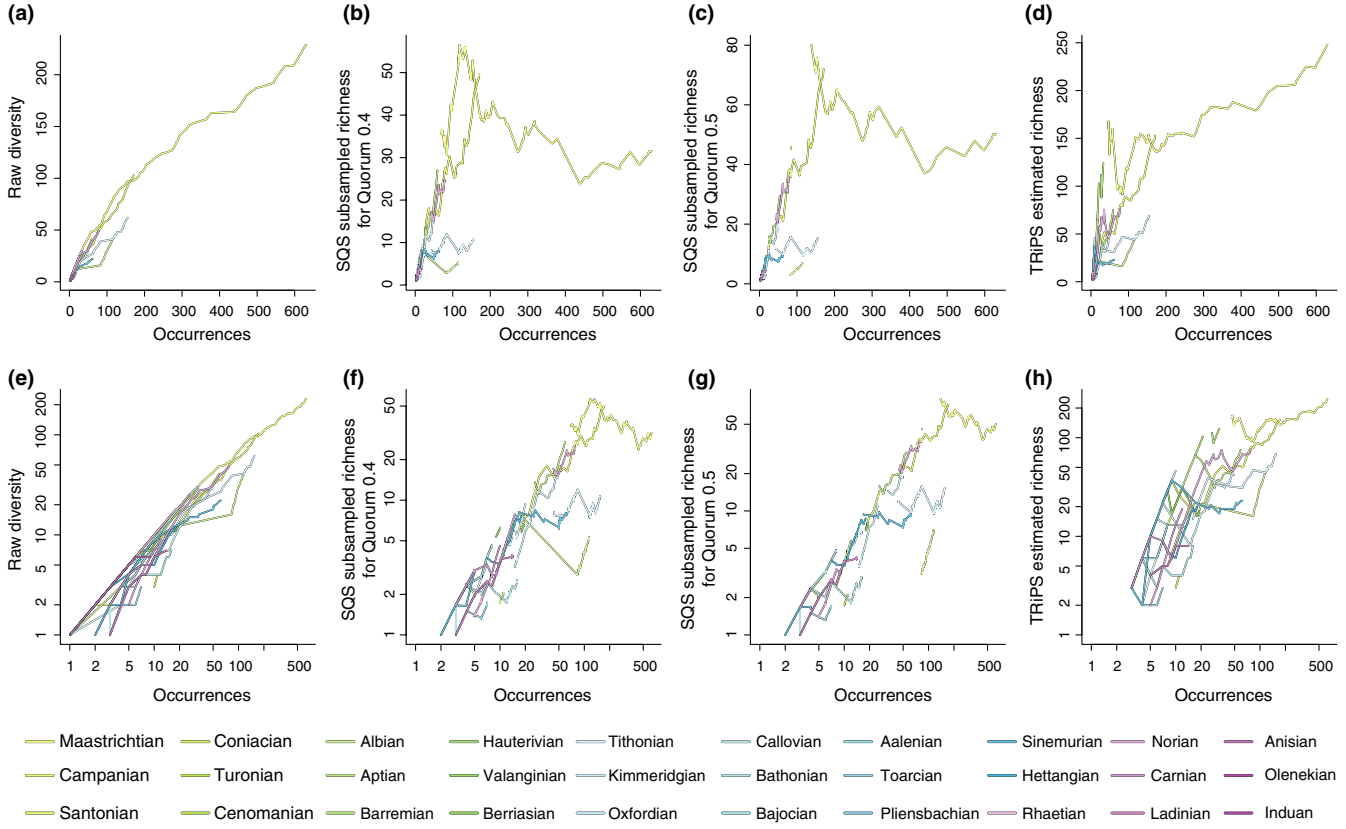


FIGURE 7 Discovery curves of global dinosaur species for each Mesozoic stage, plotted against the chronological addition of occurrences. Curves are shown using both linear and log-log axes; the latter facilitate visualisation of changes at smaller sample sizes. (a, e) Raw (unstandardised), (b, f) SQS (quorum = 0.4), (c, g) SQS (quorum = 0.5), and (d, h) TRiPS sampling-standardised species-level collector-curves

S12c): a late, steep rise is evident at a quorum of 0.6, causing the empirical curve to fall below the null. However, the empirical curves overlap with the null for quorums of 0.1–0.5. This may be because later collecting efforts do little to alter the species sampled at low quorum levels (the most common taxa). Diversity curves constructed using all the estimators we test are shown in Figures S14 and S15.

4 | DISCUSSION

Both our simulations (Figures 1, S3 and 2) and SSDCs for fossil datasets (Figures S9 and 6) demonstrate that although interpolators consistently standardise diversity samples of differing sizes, extrapolators should no more be expected to yield fair results from such samples than direct comparisons of face-value counts of species (unless coverage for all assemblages is sufficient for extrapolators to have reached an asymptote). Extrapolators yield a minimum bound for true richness (Chao, 1984). However, true richness may substantially exceed this when sample sizes are insufficient (Chao, Colwell, Lin, & Gotelli, 2009). The sample-size dependency of extrapolators is well-known (Chao et al., 2009; Colwell & Coddington, 1994) but perhaps not widely appreciated.

By their very nature, richness estimates obtained from SQS and CR are relatively sample-size invariant, and our empirical SSDC results broadly reflect this. However, extrapolated SSDCs for empirical fossil data show few signs of asymptoting (Figures 6 and S9). This suggests that sampling in the tetrapod fossil record is generally not yet good enough to use extrapolators unless they are applied within a rarefy-and-extrapolate protocol of the kind used in our simulation experiments. Our simulations show that rarefying Chao1 or λ^5 estimates to equal coverage produces the best results (particularly λ^5; Figures 5 and S5). In particular, rarefying extrapolators to equal coverage is the best way to remove confounding effects of among-sample variation in evenness, a problem that affects all richness estimators when sampling is comparatively limited (see below). A similar approach is advocated by Colwell et al. (2012) and Chao and Jost (2012). iNEXT (Hsieh et al., 2016) implements this for Chao1, but analytical solutions for adjusting λ^5 to particular levels of coverage do not yet exist.

4.1 | TRiPS

Our results suggest that, as additional data is accumulated, TRiPS eventually stops extrapolating when evenness is less than perfect because it fits a single sampling rate for all species in each interval (="uniform" sampling rates in the parlance of Wagner & Marcot,

2013). This parameter can be unduly influenced by common species: one species observed numerous times can vastly inflate the sampling rate. This appears to cause TRiPS' binomial sampling probability to saturate before coverage is complete, and thus cease extrapolating even when many species remain undiscovered. This may be seen in the records of Maastrichtian and Ypresian tetrapods (Figure S9j,c): despite incomplete sample coverage, TRiPS' richness estimates increase in step with face-value counts of species. Per-lineage binomial sampling probabilities may not have a simple relationship with frequency-distribution coverage.

Attempting to fit a single sampling rate and probability to all taxa in each sampling unit is unlikely to work on real-world data, because empirical species-abundance distributions tend to be heavily right-skewed on an arithmetic scale (Preston, 1962a, 1962b). Modern species-abundance distributions are best described by double-geometric distributions, but the lognormal is a reasonable alternative (Alroy, 2015). Although time-averaging could potentially alter these patterns, log-normal distributions of per-collection sampling rates among taxa have been shown to fit empirical fossil occurrence data much better than uniform rates (Wagner & Marcot, 2013). Even if the per-individual chance of preservation were identical for all species, ubiquitously right-skewed abundance distributions cause sampling rates and probabilities to be overestimated for rare taxa.

4.2 | Reporting biases and sampling-universe variability

Sampling-standardised discovery curves are valuable because the fossil record is not sampled in an unbiased and random manner, and because the nature of sampling may change through collector time. SQS SSDCs are considerably more stable than those for extrapolators. However, in some instances SQS richness may rise or fall as data accrues. We attribute such fluctuations to two drivers: (1) non-random reporting of fossil discoveries, and (2) sporadic expansion of the sampling universe through collector-time.

A key assumption of any richness estimator is that sampling is unbiased. However, palaeontological research probably exhibits a 'novelty bias'—a tendency to prioritise publication of new taxa over new occurrences of named taxa (Alroy, 2010c; Tennant et al., 2018). At least in the early phases of discovery, this bias results in inflated counts of singletons, which bias estimates of sample coverage downwards, and estimated richness upwards. When novel taxa become scarce, efforts may shift towards reporting additional occurrences of named taxa. This phenomenon may explain the decline in SQS richness of Maastrichtian tetrapods over the latter half of the twentieth century (see Figure S9a). The non-random nature of palaeontological reporting practices is underscored by the trajectories of singleton and multiton taxa through collector-time (Figure S9v–x). When sampling is entirely random (Figure 8), the ratio of singleton to multiton taxa is expected to decline more or less monotonically, but seems to be invariant for Maastrichtian tetrapods: multitons are underreported relative to singletons. This may explain why Good's *u* often appears to asymptote well below 1 (Figure 6s–u).

SQS SSDCs may also fluctuate due to non-random expansion of the sampling universe (e.g. increases in sampled geographic area, palaeolatitudes or palaeoenvironments). Studies of regional-level diversity patterns (i.e. continental-scale or gamma diversity) implicitly assume that fossil discoveries are a representative, random sample of that geographic region. However, fossil discoveries within continental regions have highly non-random spatial distributions, providing only a partial window into the intended geographic sampling universe. Furthermore, the realised sampling universe tends to expand as new fossiliferous regions are discovered. Even the best richness estimators cannot correct for variability in the size of the underlying taxon pool. It is, therefore, important that the realised sampling universes within focal assemblages are comparable.

SSDCs provide valuable context for gauging the maturity of sampling in focal assemblages. At the regional level, very few intervals of the dinosaur record have emerged from an early phase of discovery that tends to be characterised by volatile SQS SSDCs (Figure 7b–c and f–g). The Maastrichtian, Kimmeridgian and Tithonian of North America are likely exceptions (Figure 6). However, even the apparent stability of SQS richness in these intervals could change if productive new fossiliferous regions are discovered. This emphasises the need to recognise potential disconnects between the extent of the intended and realised sampling universes and tailor comparisons of diversity accordingly (Close, Benson, Upchurch, & Butler, 2017).

4.3 | Among-sample variation in evenness

SQS has recently been criticised for tracking evenness (Hannisdal, Haaga, Reitan, Diego, & Liow, 2017). In fact, among-sample variation in evenness will confound *any* richness estimator that implicitly or explicitly utilises information about relative frequencies of taxa (see also Kosnik & Wagner, 2006). This is simply because it becomes much harder to sample all of the species in a community when evenness is very low. We consider that any additional sensitivity SQS may have to differences in evenness at low coverage is a worthwhile tradeoff (Figures 3 and 4). The initial description of SQS (Alroy, 2010a) acknowledged the potential for among-sample variation in evenness to confound SQS; indeed, the central assumption of SQS is that substitutions of taxa occur randomly with respect to their relative frequencies. In other words, SQS is only guaranteed to estimate richness ratios with perfect accuracy when evenness (or the shape of the species abundance distribution more generally) does not vary systematically between communities.

In fact, depending on the level of sampling, standardising to equal coverage is either more *or* less sensitive to evenness compared to methods that standardise to sample size (e.g. CR). When coverage is poor to moderate, richness estimates standardised to equal coverage are marginally more sensitive to evenness than those standardised to sample size. This is because SQS establishes how many species will be found, on average, by repeatedly sampling a fixed proportion of individuals in the community. We naturally expect to sample fewer species in a given fraction of the community if evenness is very low, and more species if evenness is very high.

This is why, when sampling is comparatively limited, sample coverage (both true and estimated) actually increases as abundance distributions become more uneven (Figure S16; note that this shows the coverage *deficit* in order to allow log-transformation of the *y*-axis). When evenness is very low, coverage at smaller sample sizes is relatively high (and the coverage deficit is therefore low): although very rare species are unlikely to be sampled even once, common species are easy to find, and they collectively account for a large fraction of individuals in the population. Conversely, when evenness is very high, coverage drops, because limited samples likely contain many singletons. However, because coverage increases and asymptotes more rapidly in very even communities than very uneven ones, this relationship reverses when sampling is very good. Eventually, coverage for a given sample size will be higher when communities are more even, and lower if they are less even (Figures S16 and S6). Thus, as coverage increases, problems arising from among-sample differences in evenness diminish and eventually disappear. The implication of this changing relationship between coverage, evenness and sample sizes is that SQS is more sensitive than CR to differences in evenness at low quorum levels, because it undersamples (relative to total species richness) when evenness is low. Conversely, SQS is less sensitive than CR to differences in evenness at very high levels of coverage, because SQS then samples harder when evenness is low.

From a theoretical perspective, total species richness and the shape of the abundance distribution are distinct properties. However, practicalities of sampling mean that it may be difficult to disentangle these two properties. As Chao and Jost (2012) observe, variation in the shape of the abundance distribution is the reason why size-based (CR) rarefaction curves for different assemblages can cross (signifying points where the rank-order richness of communities switches). Coverage-based rarefaction curves (plots of richness as a function of coverage; e.g. Figure 2) cross the same number of times as size-based rarefaction curves, but less data is required to detect where this occurs. The only way to correctly resolve true differences in ranked richness is by attaining sufficient coverage in each assemblage to have observed all the crossing points—but in reality, we can never know if we have surpassed this point (Chao & Jost, 2012). This is the main reason for using the highest quorum level possible, and for treating estimates from low quorum levels with scepticism. However, SQS does tell you how many species will be found, on average, in a random sample of a fixed percentage of individuals drawn from the population, information that is biologically meaningful.

Another reason for preferring higher quorum levels—even if the shape of the abundance distribution does not vary between communities—is that it is difficult to accurately estimate low levels of coverage from limited samples. All else being equal, SQS requires much less data than CR to accurately reconstruct richness ratios and, in theory, richness ratios can be accurately reconstructed from very small sample sizes provided that abundance distributions do not differ (Chao & Jost, 2012, Table 2). In practice, however, sample coverage must be estimated from the data. Our simulations demonstrate that coverage can be very accurately estimated when sample sizes are moderately large; precision increases with sample size (Figures S17 and S18). However, both accuracy and precision depend on true richness and evenness: coverage is more difficult to estimate from small samples when true richness and evenness are low, and easier to estimate when richness and evenness are high.

5 | CONCLUSIONS

Simulations and empirical sampling-standardised discovery curves (SSDCs) for fossil datasets show that standardising diversity data to equal coverage ensures fair comparisons of richness when sampling is limited. When sampling is unbiased and the shape of the abundance distribution does not vary among communities, SQS yields perfectly accurate relative richness ratios, and standardised estimates scale linearly with true richness. Empirical SSDCs using SQS are more stable than those using extrapolators. Richness estimators that standardise by coverage are among the best currently available methods for reconstructing deep-time biodiversity patterns.

Extrapolated richness estimates obtained from samples of unequal sizes may be almost as misleading as direct comparisons of unstandardised richness. Unless sampling is sufficiently complete for the estimator to have asymptoted, extrapolated estimates may strongly depend on sample size, yielding inaccurate relative richness ratios among assemblages. This is especially crucial for fossil occurrence data, because sample completeness varies substantially among time intervals and geographic regions. The sampling level required for extrapolators to asymptote increases with true richness and decreases with evenness. Of the extrapolators we tested, the λ^5 method performs best when evenness is low.

When abundance distributions are less than perfectly even and sampling is moderate to good but not complete, TRiPS stops extrapolating and instead returns face-value counts of taxa. This is because TRiPS fits a single sampling rate for all species in each interval, which causes the method to overestimate binomial sampling probabilities. Most assemblages of interest to palaeobiologists or ecologists are unlikely to have flat abundance distributions, and indeed SSDCs using TRiPS often closely track unstandardised discovery curves.

All richness estimators are biased by differences in evenness when sampling is comparatively limited. Richness estimates become downwardly biased as evenness diminishes, since it becomes ever harder to detect the rarest species. When overall sampling is very poor, standardising by coverage produces richness estimates that are slightly more biased by differences in evenness than standardising by sample size. However, when sampling is very good, the situation is reversed.

Rarefying extrapolated richness estimators to equal sample coverage (i.e. using a coverage-based rarefaction algorithm to standardise extrapolated, rather than face-value counts of species) gives us the best of both worlds: it makes our samples effectively a little bigger, and therefore diminishes the impact of evenness while retaining the desirable properties of SQS (e.g. a linear relationship

between true and estimated richness). Coverage-based rarefaction of extrapolators removes any potential sample-size dependency, and effectively extends the maximum coverage obtainable from limited diversity samples.

Our empirical SSDCs reveal biases in the accumulation of palaeobiological knowledge that may confound even the best richness estimators. We recommend constructing SSDCs for fossil datasets in order to shed light on these sources of bias, and to provide important historical context for understanding the reliability of present-day sampling-standardised richness estimates.

ACKNOWLEDGEMENTS

This research was funded by the European Union's Horizon 2020 research and innovation programme under grant agreement 637483 (ERC Starting Grant TERRA to R.J.B.). We thank Roger Benson and Jon Tennant for discussion, and Jostein Starrfelt and an anonymous reviewer for critical but constructive comments that improved the revised manuscript. All contributors to the Paleobiology Database are thanked for their efforts, particularly M. T. Carrano, P. D. Mannion, R. B. J. Benson, A. M. Rees, W. Kiessling, M. E. Clapham, F. T. Fursich, M. Aberhan and M. D. Uhen. This is Paleobiology Database official publication no. 307.

AUTHORS' CONTRIBUTIONS

S.E. and R.A.C. independently conceived of the idea of studying how sampling-standardised estimates in the fossil record change through collector-time. R.A.C. conceived the idea for sampling-standardised discovery curves, designed and conducted the analyses, wrote the manuscript and prepared the figures. J.A. provided critical methodological input and wrote the R code for SQS (exact algorithm) and the λ^5 method. R.J.B., S.E. and J.A. provided critical comments and edits on the manuscript.

DATA ACCESSIBILITY

R code for all analyses is available from https://doi.org/10.5281/zenodo.1167536 (Close et al., 2018), and is also available on GitHub at https://github.com/rclose/SSDCs.

ORCID

Roger A. Close http://orcid.org/0000-0003-3302-9902
Serjoscha W. Evers http://orcid.org/0000-0002-2393-5621
John Alroy http://orcid.org/0000-0002-9882-2111
Richard J. Butler http://orcid.org/0000-0003-2136-7541

REFERENCES

Alroy, J. (2000). Successive approximations of diversity curves: Ten more years in the library. *Geology*, *28*, 1023. https://doi.org/10.1130/0091-7613(2000)28<1023:SAODCT>2.0.CO;2

Alroy, J. (2009). A deconstruction of Sepkoski's Phanerozoic marine evolutionary faunas based on new diversity estimates. 2009 Portland GSA Annual Meeting.

Alroy, J. (2010a). The shifting balance of diversity among major marine animal groups. *Science*, *329*, 1191–1194. https://doi.org/10.1126/science.1189910

Alroy, J. (2010b). Geographical, environmental and intrinsic biotic controls on Phanerozoic marine diversification. *Palaeontology*, *53*, 1211–1235. https://doi.org/10.1111/j.1475-4983.2010.01011.x

Alroy, J. (2010c). Fair sampling of taxonomic richness and unbiased estimation of origination and extinction rates. In J. Alroy & G. Hunt (Eds.), *Quantitative methods in paleobiology* (pp. 55–80). New Haven, CT: The Paleontological Society.

Alroy, J. (2014). Accurate and precise estimates of origination and extinction rates. *Paleobiology*, *40*, 374–397. https://doi.org/10.1666/13036

Alroy, J. (2015). The shape of terrestrial abundance distributions. *Science Advances*, *1*, e1500082–e1500082.

Alroy, J. (2017). Effects of habitat disturbance on tropical forest biodiversity. *Proceedings of the National Academy of Sciences*, *16*, 201611855–16.

Alroy, J., Marshall, C. R., Bambach, R. K., Bezusko, K., Foote, M., Fürsich, F. T., … Webber, A. (2001). Effects of sampling standardization on estimates of Phanerozoic marine diversification. *Proceedings of the National Academy of Sciences*, *98*, 6261–6266. https://doi.org/10.1073/pnas.111144698

Barrett, P. M., McGowan, A. J., & Page, V. (2009). Dinosaur diversity and the rock record. *Proceedings of the Royal Society B*, *276*, 2667–2674. https://doi.org/10.1098/rspb.2009.0352

Benson, R. B. J., Butler, R. J., Alroy, J., Mannion, P. D., Carrano, M. T., & Lloyd, G. T. (2016). Near-stasis in the long-term diversification of Mesozoic tetrapods. *PLoS Biology*, *14*, e1002359. https://doi.org/10.1371/journal.pbio.1002359

Benton, M. J. (1985). Mass extinction among non-marine tetrapods. *Nature*, *316*, 811–814. https://doi.org/10.1038/316811a0

Burnham, K. P., & Overton, W. S. (1978). Estimation of the size of a closed population when capture probabilities vary among animals. *Biometrika*, *63*, 625–633. https://doi.org/10.1093/biomet/65.3.625

Butler, R. J., Benson, R. B. J., Carrano, M. T., Mannion, P. D., & Upchurch, P. (2011). Sea level, dinosaur diversity and sampling biases: Investigating the 'common cause' hypothesis in the terrestrial realm. *Proceedings of the Royal Society B*, *278*, 1165–1170. https://doi.org/10.1098/rspb.2010.1754

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. (2005). *Species estimation and applications* (2nd ed.). Hoboken, NJ: John Wiley & Sons Inc.

Chao, A., & Chiu, C. H. (2016). Nonparametric estimation and comparison of species richness. *eLS*, 1–11. https://doi.org/10.1002/9780470015902.a0026329

Chao, A., Colwell, R. K., Lin, C.-W., & Gotelli, N. J. (2009). Sufficient sampling for asymptotic minimum species richness estimators. *Ecology*, *90*, 1125–1133. https://doi.org/10.1890/07-2147.1

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

Close, R. A., Benson, R. B. J., Upchurch, P., & Butler, R. J. (2017). Controlling for the species-area effect supports constrained long-term Mesozoic terrestrial vertebrate diversification. *Nature Communications*, *8*, 15381. https://doi.org/10.1038/ncomms15381

Close, R. A. 2018. GitHub: rclose/SSDCs: Final code for Close et al. (2018). How should we estimate diversity in the fossil record? Testing

richness estimators using sampling-standardised discovery curves. *Methods in Ecology and Evolution*. (Version v1.0). *Zenodo*. https://doi.org/10.5281/zenodo.1167536

Colwell, R. K., Chao, A., Gotelli, N. J., Lin, S. Y., Mao, C. X., Chazdon, R. L., & Longino, J. T. (2012). Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology*, *5*, 3–21. 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 Series B (Biological Sciences)*, *345*, 101–118. https://doi.org/10.1098/rstb.1994.0091

Esty, W. W. (1986). The efficiency of Good's nonparametric coverage estimator. *The Annals of Statistics*, *14*, 1257–1260. https://doi.org/10.1214/aos/1176350066

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

Gotelli, N. J., & Chao, A. (2013). Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: S. A. Levin (Ed.), *Encyclopedia of biodiversity* (pp. 195–211). Amsterdam: Elsevier. https://doi.org/10.1016/B978-0-12-384719-5.00424-X

Hannisdal, B., Haaga, K. A., Reitan, T., Diego, D., & Liow, L. H. (2017). Common species link global ecosystems to climate change: Dynamical evidence in the planktonic fossil record. *Proceedings of the Royal Society B: Biological Sciences*, *284*, 20170722–20170729. https://doi.org/10.1098/rspb.2017.0722

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.

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

Kosnik, M. A., & Wagner, P. J. (2006). Effects of taxon abundance distributions on expected numbers of sampled taxa. *Evolutionary Ecology Research*, *8*, 195–211.

Peters, S. E. (2005). Geologic constraints on the macroevolutionary history of marine animals. *Proceedings of the National Academy of Sciences*, *102*, 12326–12331. https://doi.org/10.1073/pnas.0502616102

Preston, F. W. (1962a). The canonical distribution of commonness and rarity: Part I. *Ecology Letters*, *43*, 410–432. https://doi.org/10.2307/1933371

Preston, F. W. (1962b). The canonical distribution of commonness and rarity: Part II. *Ecology Letters*, *43*, 410–432. https://doi.org/10.2307/1933371

R Development Core Team. (2015) *R: a language and environment for statistical computing*. R Foundation for Statistical Computing, Vienna. Retrieved from http://www.R-project.org [accessed August 16 2015].

Raup, D. M. (1972). Taxonomic diversity during the phanerozoic. *Science*, *177*, 1065–1071. https://doi.org/10.1126/science.177.4054.1065

Sanders Jr, H. L. (1968). Marine benthic diversity: A comparative study. *The American Naturalist*, *102*, 243–282. https://doi.org/10.1086/282541

Sepkoski Jr, J. J. (1993). Ten years in the library: New data confirm paleontological patterns. *Paleobiology*, *19*, 43–51. https://doi.org/10.1017/S0094837300012306

Sepkoski, J. J. Jr, Bambach, R. K., Raup, D. M., & Valentine, J. W. (1981). Phanerozoic marine diversity and the fossil record. *Nature*, *293*, 435–437. https://doi.org/10.1038/293435a0

Smith, A. B., & McGowan, A. J. (2011). The ties linking rock and fossil records and why they are important for palaeobiodiversity studies. *Geological Society, London, Special Publications*, *358*, 1–7.

Starrfelt, J., & Liow, L. H. (2016a). How many dinosaur species were there? Fossil bias and true richness estimated using a Poisson sampling model. *Philosophical Transactions of the Royal Society of London Series B (Biological Sciences)*, *371*, 20150219. https://doi.org/10.1098/rstb.2015.0219

Starrfelt, J., & Liow, L. H. (2016b) Data from: How many dinosaur species were there? Fossil bias and true richness estimated using a Poisson sampling model. *Dryad Digital Repository*. https://doi.org/10.5061/dryad.86922

Tennant, J. P., Chiarenza, A. A., & Baron, M. (2018). How has our knowledge of dinosaur diversity through geologic time changed through research history? *PaleorXiv*, 1–35. osf.io/preprints/paleorxiv/yab83.

Upchurch, P., Mannion, P. D., Benson, R. B. J., Butler, R. J., & Carrano, M. T. (2011). Geological and anthropogenic controls on the sampling of the terrestrial fossil record: A case study from the Dinosauria. *Geological Society, London, Special Publications*, *358*, 209–240. https://doi.org/10.1144/SP358.14

Valentine, J. W. (1969). Patterns of taxonomic and ecological structure of the shelf benthos during Phanerozoic time. *Palaeontology*, *12*, 684–709.

Vavrek, M. J., & Larsson, H. C. E. (2010). Low beta diversity of Maastrichtian dinosaurs of North America. *Proceedings of the National Academy of Sciences*, *107*, 8265–8268. https://doi.org/10.1073/pnas.0913645107

Wagner, P. J., & Marcot, J. D. (2013). Modelling distributions of fossil sampling rates over time, space and taxa: Assessment and implications for macroevolutionary studies. *Methods in Ecology and Evolution*, *4*, 703–713. https://doi.org/10.1111/2041-210X.12088

SUPPORTING INFORMATION

Additional Supporting Information may be found online in the supporting information tab for this article.

How to cite this article: Close RA, Evers SW, Alroy J, Butler RJ. How should we estimate diversity in the fossil record? Testing richness estimators using sampling-standardised discovery curves. *Methods Ecol Evol*. 2018;00:1–15. https://doi.org/10.1111/2041-210X.12987