ECOGRAPHY

Research

A comparison of diversity estimators applied to a database of host–parasite associations

Claire S. Teitelbaum, Caroline R. Amoroso, Shan Huang, T. Jonathan Davies, Julie Rushmore, John M. Drake, Patrick R. Stephens, James E. Byers, Ania A. Majewska and Charles L. Nunn

C. S. Teitelbaum (https://orcid.org/0000-0001-5646-3184) ✉ (claire.teitelbaum@gmail.com), Odum School of Ecology, Athens, GA, USA. – C. R. Amoroso (https://orcid.org/0000-0001-5456-7975). – S. Huang (https://orcid.org/0000-0002-5055-1308). – T. J. Davies (https://orcid.org/0000-0003-3318-5948). – J. Rushmore (https://orcid.org/0000-0002-2682-6893). – J. M. Drake (https://orcid.org/0000-0003-4646-1235). – P. R. Stephens (https://orcid.org/0000-0003-1995-5715). – J. E. Byers (https://orcid.org/0000-0001-9240-0287). – A. A. Majewska (https://orcid.org/0000-0002-0965-3177). – C. L. Nunn (https://orcid.org/0000-0001-9330-2873).

Understanding the drivers of biodiversity is important for forecasting changes in the distribution of life on earth. However, most studies of biodiversity are limited by uneven sampling effort, with some regions or taxa better sampled than others. Numerous methods have been developed to account for differences in sampling effort, but most methods were developed for systematic surveys in which all study units are sampled using the same design and assemblages are sampled randomly. Databases compiled from multiple sources, such as from the literature, often violate these assumptions because they are composed of studies that vary widely in their goals and methods. Here, we compared the performance of several popular methods for estimating parasite diversity based on a large and widely used parasite database, the Global Mammal Parasite Database (GMPD). We created artificial datasets of host–parasite interactions based on the structure of the GMPD, then used these datasets to evaluate which methods best control for differential sampling effort. We evaluated the precision and bias of seven methods, including species accumulation and nonparametric diversity estimators, compared to analyzing the raw data without controlling for sampling variation. We find that nonparametric estimators, and particularly the Chao2 and second-order jackknife estimators, perform better than other methods. However, these estimators still perform poorly relative to systematic sampling, and effect sizes should be interpreted with caution because they tend to be lower than actual effect sizes. Overall, these estimators are more effective in comparative studies than for producing true estimates of diversity. We make recommendations for future sampling strategies and statistical methods that would improve estimates of global parasite diversity.

Keywords: biodiversity, comparative analysis, databases, parasite species richness, sampling bias, simulations

Introduction

Understanding the worldwide distribution and diversity of species is important for conserving biodiversity in the face of anthropogenic environmental change, yet our knowledge of the total diversity of organisms is limited (May 1988, Mora et al. 2011, Edie et al. 2017, Pappalardo et al. 2020). The diversity of parasitic organisms is particularly poorly quantified, even though parasites are believed to make up over half of the total species on the planet (Poulin and Morand 2000, Dobson et al. 2008, Carlson et al. 2019). Host–parasite interactions are ecologically important, but attempts to quantify total parasite diversity and understand its drivers can be undermined by uneven sampling across space and taxa.

Sampling of parasites is unevenly distributed among host and parasite taxa (Cooper and Nunn 2013, Stephens et al. 2016, 2017). Hosts and parasites with certain traits are more likely to be sampled than others. For example, primates are more likely to be sampled if they are ground-dwelling than if they are arboreal (Cooper and Nunn 2013). Because hosts are a fundamental part of parasites' environments, many studies have attempted to identify drivers of parasite diversity among host lineages (Nunn et al. 2003, 2005, Bordes et al. 2007, Huang et al. 2015, Teitelbaum et al. 2018). Differential sampling effort can confound these comparative studies because host species that have received more research also have more parasites reported (Walther et al. 1995, Blakeslee and Byers 2008). Further, variation in sampling effort often covaries with predictors of parasitism, such as geographic range area (Gregory 1990), body mass, and movement patterns (Teitelbaum et al. 2018). Thus, comparative studies that seek to identify relationships between parasite diversity and biological traits of host species need to account for uneven sampling effort across hosts (Walther et al. 1995, Jorge and Poulin 2018). Studies of large-scale variation in parasite diversity have used various strategies to account for heterogeneity in sampling (Box 1). Assessing the performance of these methods is challenging, however, because the true diversity of parasites is unknown (but see assessments based on systematic surveys by Poulin 1998, Walther et al. 2005).

Recently-compiled large databases of host–parasite associations provide an opportunity to study patterns of parasite diversity across ever-larger scales, but also highlight the substantial variation in sampling effort across taxa and geography.

Box 1. Methods for controlling for sampling effort and their assumptions

The methods commonly used to control for sampling effort fall into three broad categories: methods with no estimation, species accumulation (rarefaction), and nonparametric estimators. All of the methods we consider in this study are incidence-based, meaning that they use presence/absence rather than abundance data. Most assume that the number and/or frequencies of detected species somehow represent the number of undetected species (Gotelli and Colwell 2011). Similarly, rarefaction and nonparametric estimators assume that undetected species are missed randomly rather than systematically (Walther and Moore 2005). Last, all methods assume that the total number of species (and their abundances) is constant (Gotelli and Colwell 2011).

Methods with no estimation do not consider the differential likelihood of different parasite species being included in the database. Raw sampling does not control for differential sampling effort, but rather assumes that raw sampled parasite species richness (PSR) accurately represents true PSR. This method is rarely used except when sampling is exhaustive (Walther and Morand 1998, Jorge and Poulin 2018), but provides a baseline for minimum true PSR. Dataset reduction uses raw sampling combined with a citation cutoff (i.e. reducing the sample to only the host species with at least a minimum number of citations in the database) (Cooper et al. 2012). Citation residuals uses residuals from a regression of raw richness against a measurement of sampling effort (e.g. number of citations per host species) to control for differential sampling effort (Nunn et al. 2003) or includes a measurement of sampling effort as a covariate in a linear model (Altizer et al. 2007).

Species accumulation methods make no assumptions about the relative abundances of species within a community and use resampling from the raw data to compare diversity among sampled assemblages or to estimate asymptotic species richness. Rarefaction accounts for differential sampling effort across host species by interpolating data back to a common sampling effort (e.g. the lowest present in the dataset). Rarefaction thus does not estimate true diversity, but rather attempts to create a dataset that scales with true diversity. The use of rarefaction requires comparable sampling methods across samples, nestedness of samples, random placement of individuals and independent samples (Sanders 1968, Blakeslee and Byers 2008, Gotelli and Colwell 2011).

In contrast to the interpolation of rarefaction, non-parametric methods extrapolate the dataset to estimate true PSR and may be effective with smaller sample sizes. Chao2 uses a ratio of singletons (for our purposes, parasite species found in only one study of a host species) to doubletons (parasite species found in two or more studies of a host) to calculate an estimated richness. This calculation assumes that a high ratio of singletons to doubletons implies that there are still many species remaining to be found, and as doubletons become more common, fewer species remain undiscovered. Following a similar logic, the Incidence Coverage Estimator (ICE) estimates the true richness using a ratio of frequently and infrequently occurring species. The jackknife estimate is based on how many species are lost when samples (studies) are removed, using the total number of samples and the number of singletons (for first-order jackknife) or the number of singletons and doubletons (second-order jackknife). The bootstrap method involves a calculation of the species present and absent in a random resampling draw from a distribution of samples. Methods for calculating each of these estimates are described in detail in Gotelli and Colwell (2011).

These databases span multiple taxonomic groups, often at a global scale (reviewed by Stephens et al. 2016). In these databases, a single host species could be represented by hundreds of published studies conducted over broad geographic and temporal ranges, collectively reporting hundreds of parasites infecting a single host species. At the other extreme, some host species are represented by only a single published study, which likely reports a limited selection of parasite species at a single location and time. Between these extremes, other species exhibit variation in sampling effort, and this variation drives much of the apparent variation in parasite species richness across host species (Nunn et al. 2003, Stephens et al. 2016).

Another source of variability in compiled databases is that the primary studies that constitute these databases are conducted with a variety of goals and methods. The published literature that provides the building blocks of large databases of parasite diversity (e.g. Berger 2005, Gibson et al. 2005, Yu and Edberg 2005) ranges from targeted studies focusing on a single parasite species to broad, exploratory surveys aimed at identifying entire communities of parasites in natural systems (Nunn and Altizer 2006, Olival et al. 2017). As a result, sampling is biased toward some parasites over others (e.g. those relevant to human health), both within and across host species. Methods of controlling for differential sampling effort commonly assume that each study randomly samples the full community of species, yet databases compiled from sources with diverse goals and methodologies often violate this assumption (Lobo et al. 2018). The number of parasite species considered in a single study can range from one in studies typically targeting a disease of particular interest (e.g. Acevedo-Whitehouse et al. 2003) to tens of parasite species in a comprehensive search (e.g. Schaffer et al.'s (1981) study of raccoon helminths). Thus, while these large databases provide an opportunity to investigate broad-scale drivers of parasite diversity, they also pose the challenge of controlling for multiple components of sampling bias. It is necessary to assess how effectively existing methods estimate parasite diversity for these types of datasets, and to consider how future sampling in primary studies could be targeted to address sampling bias at its source.

Here, we combine a global database of parasite occurrences with simulations to 1) evaluate the performance (i.e. precision and bias, Walther and Moore 2005) of commonly used methods for estimating parasite diversity given incomplete and uneven sampling (Box 1), 2) evaluate modeling methods for controlling for sampling effort in comparative studies and 3) inform strategies for future sampling of parasite diversity. These methods have been shown to be useful for systems that have been systematically sampled (Walther and Morand 1998, Chao et al. 2009, Gotelli and Colwell 2011), but have rarely (if ever) been assessed in databases with such large variation in sampling effort and goals. For 3), we perform simulations to identify the types of primary field studies that, if added to the database, would most efficiently improve the bias and precision of these estimators.

Methods

The empirical database

Our simulations are based on the Global Mammal Parasite Database (GMPD) (Stephens et al. 2017). The GMPD contains records of infection in free-ranging populations of four mammalian orders: Artiodactyla, Carnivora, Perissodactyla and Primates (excluding humans) across a broad range of parasites (helminths, arthropods, bacteria, fungi, protozoa and viruses). Each record in the GMPD represents a unique host–parasite-study combination; some records contain additional information such as prevalence and number of individuals sampled. We excluded records from the GMPD where the host and/or parasite was not identified to the genus or species level or where the recorded prevalence of the parasite was zero (i.e. known absences); we thus included presence-only data in this analysis. We also included records where the prevalence was not reported. The resulting database includes 1996 parasite species identified in 461 host species, reported in 2632 published studies.

When counting the number of parasite species identified in each host species (i.e. parasite species richness, PSR), we included parasites identified to the genus level only if no specific parasite of that genus had been identified in the same host species (e.g. a record of *Salmonella* sp. was counted only if no record of *Salmonella* in the same host species had been identified to the species level). In the current version of GMPD, the recorded PSR for any given host species ranges from 1 (for 73 host species) to 159 (in raccoons *Procyon lotor*).

We quantified the sampling effort for a given host species as the number of unique studies in the database that sampled that host species. One alternative method for quantifying sampling effort is to count the total number of individuals that have been sampled across studies for a given host species. We elected to measure sampling effort using the number of studies because not all studies (86%) in our dataset reported the number of individuals sampled, and citation counts more mechanistically represent the research biases that are relevant to databases (e.g. two studies that take different approaches to sampling provide more information than a single study that samples twice as many individuals for the same parasite species). Further, the two approaches tend to produce congruent results in comparative studies, likely because there is a strong relationship between the number of studies and the cumulative number of individuals sampled per host species (e.g. in the GMPD, r = 0.75) (Gregory 1990, Nunn et al. 2003, Ezenwa et al. 2006). Therefore, we focus our effort on simulating the number of studies, with the goal of providing a general framework that can be applied to additional measures of sampling effort in other databases. It is also important to note that this metric is distinct from the number of studies per host species in the scientific literature in general (e.g. number of citations in Web of Science), which is another metric employed in comparative studies (Nunn et al. 2003, Teitelbaum et al. 2018).

Simulation of artificial datasets

We simulated host–parasite associations and their sampling to create artificial databases in which we assigned a true PSR to each host. True PSR is not known for any species, because we cannot be sure that the parasite community of any host – even the best-studied – has been completely sampled. Thus, we simulated true PSR by sampling from a simulated database according to parameters derived from the GMPD and other sources. We compared several different plausible distributions of true PSR, constrained by upper and lower limits based on the literature. We considered four possible maximum values of PSR (PSR_{max}): 800, 500, 300 and 150 (Table 1). We based these numbers on estimates of human parasites, which range from at least 1400 (Cleaveland et al. 2001) to as many as 2107 known species of parasites (Dunn et al. 2017). We expect that wildlife species have fewer parasites than humans because of their smaller geographic ranges, lower population densities and lower recorded parasite diversity (Cleaveland et al. 2001, Kamiya et al. 2014a). The lower values we used represent the scenario in which current sampling has approached true diversity in the best-sampled species. We compared two possible minimum values of PSR (10 and 50). We also examined three different beta-distribution shapes of PSR across hosts: a uniform distribution (i.e. hosts equally likely to have any PSR within the range), a right-skewed distribution (i.e. most hosts have relatively low PSR), and a centered distribution (i.e. most hosts have intermediate PSR) (Fig. 1, Table 1). In total, we compared 24 probability distributions of true PSR in our analyses.

For each of the 461 hypothetical host species, we simulated a true PSR and sampling of this true parasite diversity in four steps. In step 1, we drew the true PSR of a host (p) from one of the 24 distributions described above. This number represented parasite diversity, composed of p parasites whose taxonomic identity was not defined (Fig. 1). We assumed that host's parasite community was composed of the p most-commonly sampled parasites in the GMPD, with taxonomic identities removed.

We then simulated processes of sampling this true parasite diversity based on patterns present in the GMPD. In step 2, we assigned a number of studies, c, to the host species from the frequency distribution of study counts among host species in the GMPD, which ranged from 1 to 218 (Fig. 1).

In step 3, to each of these c studies, we assigned a number of observed parasite species, n. We drew n from the frequency distribution of the number of parasites per study across the whole GMPD, which ranged from 1 to 39 (Fig. 1). If the simulated true PSR of a host (p) was less than 39, we sampled only from studies that studied at most p parasites. We assumed that c and n were independent of one another and of the true PSR of the host species (as for spatial patterns in sampling effort, Soberón et al. 2007). In step 4, we randomly sampled the n parasite species in each study from the simulated true parasite species diversity of that host. This last step linked the simulated sampling process to the simulated true PSR, which allowed us to assess the performance of the different methods used for estimating true diversity.

In step 4, we performed simulations both with and without sampling bias among parasite species. In the GMPD, many parasite species are sampled in multiple studies of the same host for reasons unrelated to their abundance or spatial range (e.g. Yellow Fever Virus is frequently sampled in wild primates because of its relevance to human health). This bias could obscure our interpretation of the raw PSR, for example, when sampling intensity is included as a covariate in comparative analyses (Nunn et al 2003, Huang et al 2015) as well as inferences of true PSR using incidence-based methods (Box 1). In simulations that included sampling bias among parasite species, we first counted the number of studies per parasite species across the entire GMPD as a metric of relative sampling effort for each parasite. For each study in our simulations, we then used this distribution to weight the n draws of parasite identities, where the probability of sampling parasite 'species' P ($P \in \{1, 2, \ldots p\}$) was proportional to the number of citations for rank-ordered parasite species P in the GMPD. Draws were without replacement within a single study (i.e. each parasite species could only be sampled once per study), but a parasite species could be sampled in multiple studies of the same host. In simulations without sampling bias, we assumed that each parasite species was equally likely to be sampled, so we drew identities of each parasite in each study from a uniform distribution.

In addition to this global analysis, we performed simulations for subsets of hosts and parasites (Table 1) to examine how differences in sampling structure among major host lineages and parasite groups affected estimator performance. We also implemented simulations with a citation cutoff that

Table 1. Parameters used in simulation and analysis of sampling datasets. Values in bold are parameters used in simulations evaluating the efficacy of different methods to control for sampling in a comparative study and in the text when they are not otherwise mentioned (see 'Analysis 2: using estimators for comparative studies').

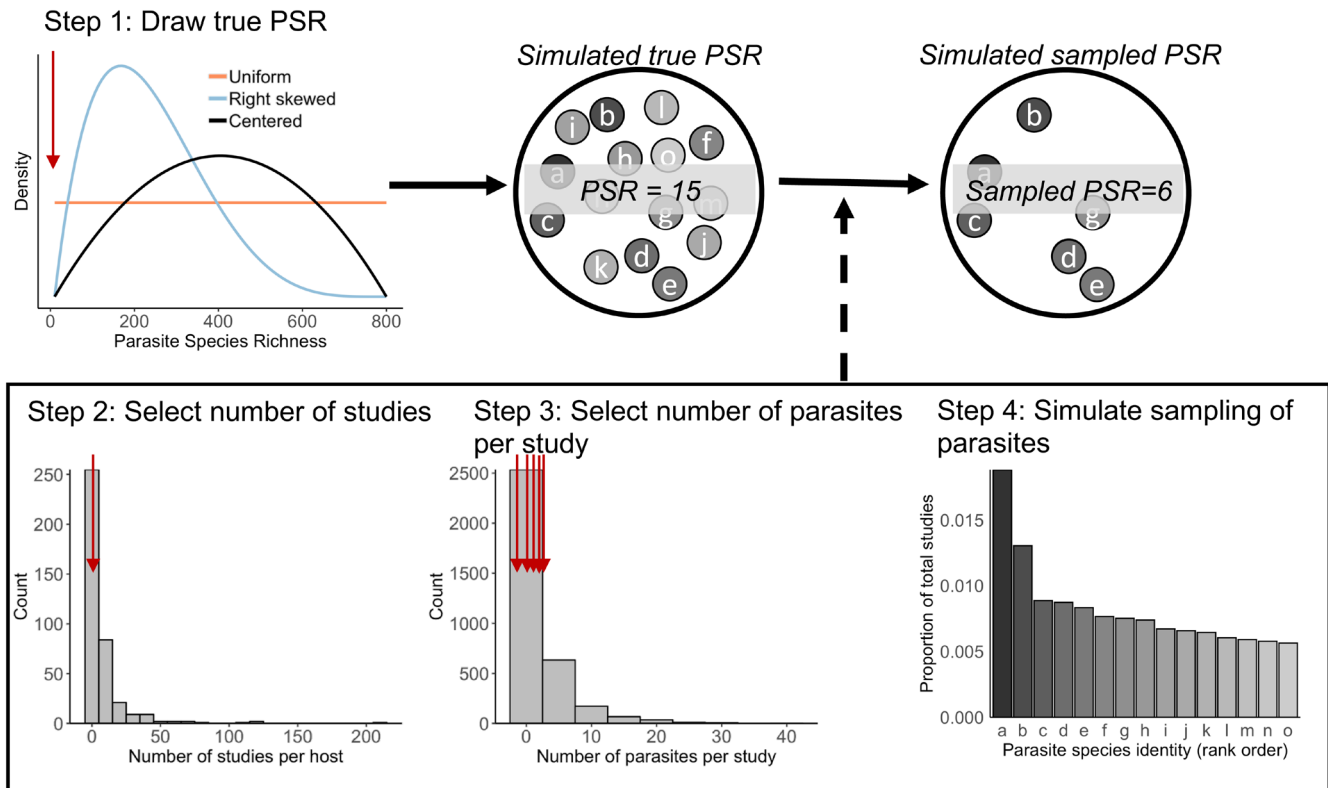| Parameter	Values
Maximum true PSR (PSR_{max})	150, 300, **500**, 800
Minimum true PSR	**10**, 50
Distribution of PSR	Uniform; centered (beta distribution with $\alpha=2$, $\beta=2$); **right-skewed (beta distribution with $\alpha=2$, $\beta=5$)**
Sampling bias among parasite species	**Yes**, no
Host group	**All**; primates only; carnivores only; ungulates only
Parasite group	**All**; helminths only; viruses only; bacteria only; protozoa only
Minimum citation cutoff	**2**, 5, 10, 15, 20

Figure 1. Simulation procedures. Histograms represent parameters based on distributions in the GMPD and density (line) plots represent hypothetical distributions. Arrows represent values selected for one example simulation. In step 1, we randomly selected a value of true PSR (*p*) for a host species from a distribution characterized by a minimum, maximum and shape. In steps 2–4, we simulated sampling to produce the sampled PSR from the true PSR (dashed arrow). Step 2: we drew the number of studies for this host species from the empirical distribution of studies per host (e.g. 7, shown by the arrow). Step 3: we drew the number of parasites per study from the empirical distribution of parasites per study (e.g. 1, 2, 2, 3, 3, 4 and 5 parasites for these 7 studies). Step 4: we simulated which parasites were sampled in each study by weighting the probability of sampling each parasite from the host's simulated community based on sampling rates across the entire GMPD (i.e. the number of studies per parasite species in the GMPD). A parasite species studied many times is more likely to be chosen than a parasite recorded in a single study. This distribution is based on the top *p* rank-ordered parasite species in the GMPD, where *p* is the host's simulated true PSR. We also preformed simulations without sampling bias, where parasite identities were drawn from a uniform distribution.

reduced the dataset to hosts with at least a minimum number of citations (Table 1) to test the effects of dataset reduction on estimator performance.

We repeated this entire process 50 times for each parameter set (Table 1), producing 100 800 simulated datasets that reflected the GMPD's apparent sampling structure. Each entry in a simulated dataset represented a host–parasite-study association. For our purposes, we did not make assumptions about parasite sharing across host species, relatedness among parasites or hosts, or potential connections between true PSR and sampling intensity, but we discuss their relevance and provide suggestions for further investigations.

Analysis 1a: estimating true diversity

In this analysis, we evaluated the precision and bias (hereafter, performance) of seven methods that are commonly used to control for differential sampling effort in diversity studies (rarefaction, citation residuals, and five biodiversity estimators: Box 1) to determine their value for studies that aim to quantify true parasite diversity. For rarefaction, we calculated a sampled PSR for each host that was rarefied down to a consistent level of sampling across all hosts, which we set at the smallest number of citations for a single host in the simulated dataset. If this number was lower than five, we included only hosts with at least five citations. We also obtained residuals from a regression of raw PSR against number of citations (Nunn et al. 2003). The biodiversity estimators were Chao2 (Chao 1984), first-order and second-order jackknife (Burnham and Overton 1978), the Incidence-based Coverage Estimator (ICE, Chao and Yang 1993), and bootstrap (Smith and van Belle 1984). For ICE, we used k = 10 as the cut-off point that separates species into 'frequent' and 'infrequent' (Chao et al. 2016). We used the sampled PSR as a null comparison to examine the performance of estimates of PSR relative to making no correction for sampling effort.

We compared the performance of the different estimators using linear models relating estimated PSR (the independent

variable) to the true PSR (the dependent variable) in simulated databases. We quantified bias of the estimates in three ways: the estimated slope, the estimated intercept, and the difference between the observed slope of the regression and the slope of a regression forced through the origin (Brose et al. 2003, Walther and Moore 2005). We used the R^2 value to quantify precision (Brose et al. 2003, Walther and Moore 2005). Unbiased estimates of PSR would have a fitted slope of 1 with an intercept of 0, and precise estimates would have R^2 values close to 1.

To investigate how the performance of each of these seven methods depended on sampling patterns and true PSR, we used linear models to assess estimator precision (i.e. the R^2 value mentioned above) as a function of the estimator used, PSR_{max}, citation cutoff, and presence of sampling bias among parasites (Table 1). We also included all pairwise interactions between these four variables. We used two citations as our lowest value for citation cutoffs because some estimators require a sample size of at least two. For these models we held other parameters constant, which were: all host groups, all parasite groups, minimum true PSR of 10, and a right-skewed distribution of true PSR. We used the right-skewed distribution based on the finding that high parasite loads (and thus, potentially high PSR) tend to occur in a relatively small proportion of hosts (Woolhouse et al. 1997, Shaw et al. 1998).

Analysis 1b: using estimators in combination with linear models for comparative studies

Many studies are not concerned with accurately estimating raw parasite diversity, but rather attempt to understand associations between PSR and biological traits of host species (e.g. body mass, geographic range size and host behavior) (reviewed in Kamiya et al. 2014a). In these types of studies, linear modeling methods can further account for sampling effort and enable comparative studies to recover true relationships between PSR and host traits. To evaluate these modeling methods, we simulated data to represent a linear relationship between a hypothetical host trait and the true PSR among host species:

$$PSR_i = \alpha + \beta \text{ trait} + \varepsilon_i$$

where PSR_i is the true PSR of species i. For simplicity, we modeled a relationship where $\alpha = 0$ and $\beta = 1$. In all cases, ε was normally distributed with a standard deviation that ranged from 1 to 500 (for comparison, the standard deviation of raw PSR in the dataset is 230) (Supplementary material Appendix 1 Fig. A1). This increasing error is designed to represent other unmeasured variables that affect PSR and to effectively decrease the effect size of the true relationship.

Because the Chao2 estimator was one of the top-performing estimators in estimating PSR (see Results), we focused on comparing Chao2 with raw sampled PSR in these analyses, using simulated datasets for one focal parameter set (Table 1). We also examined how ε affected the accuracy of estimating

β using raw or Chao2-estimated PSR. To test for type I error, we also simulated a null relationship in which the trait was unrelated to PSR (i.e. $\beta = 0$).

For both raw and Chao2-estimated PSR, we fitted: 1) univariate linear regressions predicting PSR (raw or estimated) from the host trait, 2) multivariate linear models that also included the number of citations (i.e. sampling effort) as a covariate and 3) weighted regression models, where the model was univariate but included the number of citations as weights. In all cases, PSR, the trait, and citations were log-transformed, following empirical studies (Ezenwa et al. 2006, Han et al. 2015).

Analysis 2: increasing sampling effort

In addition to identifying the methods that most effectively control for sampling effort in current databases, it is important to understand how additional sampling could most efficiently improve our ability to estimate true PSR. Thus, we simulated additional sampling beyond the current records in the dataset at 1) the host level (i.e. more studies per host species, step 2 above) and 2) the study level (i.e. more parasites per study, step 3 above). We considered increases of 5, 10, 20, 50, 100, 200 or 1000% of current sampling intensity at both the host and study levels.

At the host level, we simulated increased sampling in a given percentage of host species (e.g. 5% of hosts receive additional sampling). We randomly selected these hosts, either from the pool of all hosts or from the subset of hosts that had fewer than a cutoff number of citations (2, 3, 5, 10 or 15 citations). At the study level, we randomly assigned the additional samples to studies (thus to any host). To assess the effect of eliminating sampling bias among parasite species (non-targeted sampling), we used a uniform probability distribution of sampling any given parasite in a host (in comparison to the empirical distribution derived from the GMPD shown in Fig. 1).

Last, we compared results from different values of PSR_{max} to assess how the efficacy of each sampling scheme depended on range of true PSR values. As with the original simulated datasets, we used linear models to predict true PSR from estimated PSR when evaluating the effect of different sampling strategies.

All analyses were conducted in R 3.5.0 (R Development Core Team) with packages 'fossil' and 'SpadeR' for diversity estimation (Vavrek 2011, Chao et al. 2016) and package 'vegan' for rarefaction (Oksanen et al. 2018).

Results

Analysis 1a: performance of diversity estimators for estimating PSR

Sampling effort in the GMPD (i.e. the number of studies per host) explained a large degree of variation in raw PSR (in a linear model with both variables log-transformed, $R^2 = 0.66$).

Across all estimation methods and parameter sets, estimated PSR explained between 0% and 99% of the variation in true PSR. The performance of all methods across the simulated datasets improved with higher citation cutoffs, lower PSR_{max}, and in the absence of sampling bias among parasites (Fig. 2, Supplementary material Appendix 1 Table A1). For example, estimated PSR explained less than 5% of the variation in true PSR when including host species sampled in two or more studies but 29% of the variation in true PSR when including only host species sampled in at least 20 studies (see Table 1 for other parameters held constant) (Supplementary material Appendix 1 Fig. A2). The estimators were also less biased (i.e. slopes steeper and closer to 1, intercepts closer to 0, and smaller differences in slope compared to a regression fit through the origin) when only better-studied hosts were included (Supplementary material Appendix 1 Fig. A3–A5). However, introducing a citation cutoff decreased the sample size (e.g. only 57 of 389 hosts have been the subject of at
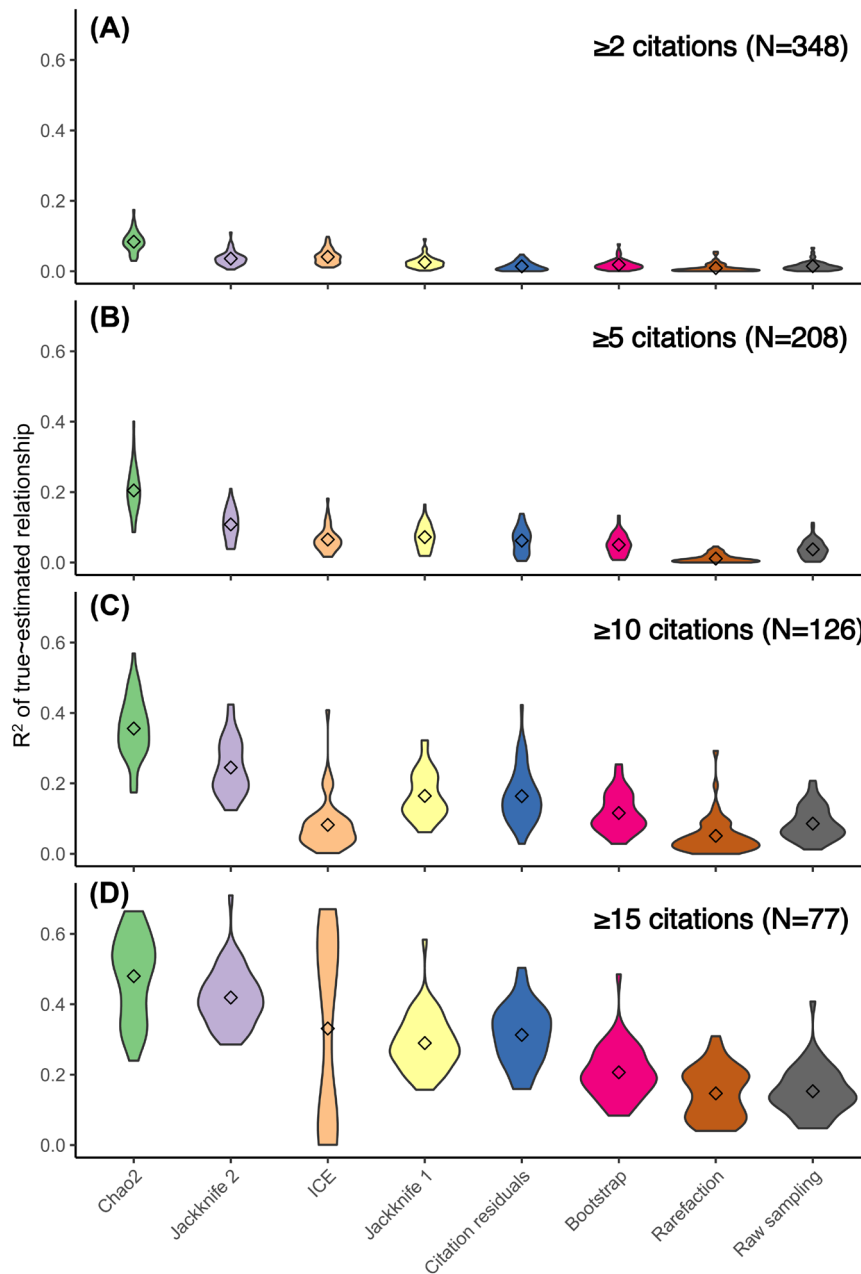

Figure 2. The precision of parasite species richness (PSR) estimates varies across methods and improves with increasing sampling intensities. Each violin represents the density distribution (and range) of precision across 50 simulations, measured by the R^2 of a linear model predicting true PSR from estimated PSR (see parameters in Table 1). Points indicate the mean. The four panels represent different subsets of hosts, selected based on sampling intensity (number of citations). Estimators are sorted by their performance across all simulated parameter combinations.

least 15 studies) (Fig. 2). When we limited the dataset to taxonomic subsets of hosts and parasites or when we used a different distribution of true PSR, results were similar to those obtained with the full dataset (Supplementary material Appendix 1 Fig. A2).

Nonparametric estimators consistently performed better than other techniques (Fig. 2, Supplementary material Appendix 1 Table A1). Differences among these estimators were small, though the Chao2, ICE and second-order jackknife were slightly more precise than others (Supplementary material Appendix 1 Table A1). Across all of our simulations, the R^2 values for the ICE were 0.058 lower, for second-order jackknife were 0.065 lower, and for first-order jackknife were 0.108 lower than for Chao2. The bootstrap estimator, citation residuals and raw sampled PSR were all substantially (> 0.11) lower than the Chao2 and second-order jackknife. Rarefaction was only marginally more precise than raw PSR. The Chao2 and ICE estimators were the least biased estimators, especially at low citation cutoffs (Supplementary material Appendix 1 Fig. A4, A5). Based on these results, we focused on the Chao2 in further analyses. Chao2 precision was correlated with per-species sampling effort (i.e. the number of studies per host species, Supplementary material Appendix 1 Fig. A6), so we also included this metric of sampling effort in our analysis on recovering a simulated correlation between PSR and a host trait (i.e. as a covariate or as model weights).

Analysis 1b: using estimators in combination with linear models for comparative studies

Across all parameter sets and modeling methods, models using Chao2 estimates of PSR recovered the simulated relationship between PSR and a trait better than models using raw PSR (Fig. 3). Using Chao2 in a weighted regression model (i.e. where predictor variables were weighted by the number of studies of parasites for a host) recovered a true relationship marginally more effectively than unweighted regression models, particularly for low citation cutoffs. The ability of all regression models to detect a significant relationship between PSR and a host trait increased as the effect size of the true relationship increased. Estimates of effect sizes were always lower than true effect sizes, regardless of the true effect size. On average, type I error rates were slightly elevated above the confidence level of $\alpha = 0.05$ (average across all cases: 0.064, confidence interval [0.050, 0.078]). Type I error rates were lowest (and averaged below 0.05) for models that used a higher citation cutoff (Supplementary material Appendix 1 Table A3).

Analysis 2: increasing sampling effort

Finally, we assessed how increased sampling effort would improve performance of the estimators. When we simulated increased sampling at the host and study levels, the performance of both the Chao2 estimator and raw sampling improved (Fig. 4). At 1000% of current sampling effort, Chao2 estimates explained up to 85% of the variation

in true PSR (when $PSR_{max} = 500$, Fig. 4A). Performance improved the most with increased sampling at the host level (i.e. more studies per host species, rather than more parasites per study). In addition, unbiased sampling schemes, where all parasite species present are equally likely to be sampled, more efficiently increased the precision of the Chao2 estimator, approaching R^2 values of 1 in some cases (Fig. 4). These estimates were also less biased if we assumed that PSR_{max} was lower, in which case sampling would capture a higher proportion of true diversity.

Discussion

Understanding broad-scale patterns of parasite diversity across host species requires accounting for differential sampling effort. We used simulated datasets to compare methods of controlling for uneven sampling effort across host species in large databases with heterogeneous sampling methods. Overall, our results highlight that it is difficult to realistically estimate PSR from such databases, which are not only limited by low levels of primary sampling but also biased towards certain hosts and parasites. Despite these challenges, nonparametric estimators, and the Chao2 and second-order jackknife estimators in particular, had the greatest potential to control for differential sampling effort across the simulated datasets with a variety of characteristics. These estimators were more effective at recovering relationships in our simulated comparative studies than in obtaining unbiased estimates of PSR. The exception to this finding was the bootstrap estimator, which performed almost as poorly as non-estimation and rarefaction-based methods (despite its high performance in some prior studies, Supplementary material Appendix 1 Table A4). Recent comparative studies using data from the GMPD have used Chao2 to account for sampling bias (Huang et al. 2015, Teitelbaum et al. 2018), and our results confirm that this method was an appropriate choice.

Previous studies at smaller spatial and taxonomic scales with unbiased sampling schemes have also found that Chao2 was among the least biased estimators for PSR (Walther and Morand 1998, Walther and Moore 2005) and that the Chao2 and jackknife estimators were the most effective for free-living species (Supplementary material Appendix 1 Table A4). Yet, this is the first time these methods have been evaluated on data with such complex variation in sampling effort, i.e. using a large database of parasite diversity with data compiled from multiple independent sources. We propose that the Chao2 and jackknife estimators might have performed better than other estimators because they better capture the unique characteristics of parasite sampling processes. Parasite diversity is generally less completely sampled than the diversity of free-living species (Carlson et al. 2019); thus, the methods that count singleton and doubleton species of parasites (i.e. Chao2, first-order jackknife and second-order jackknife) could be more effective than resampling the few studies present (e.g. as in rarefaction and bootstrapping)
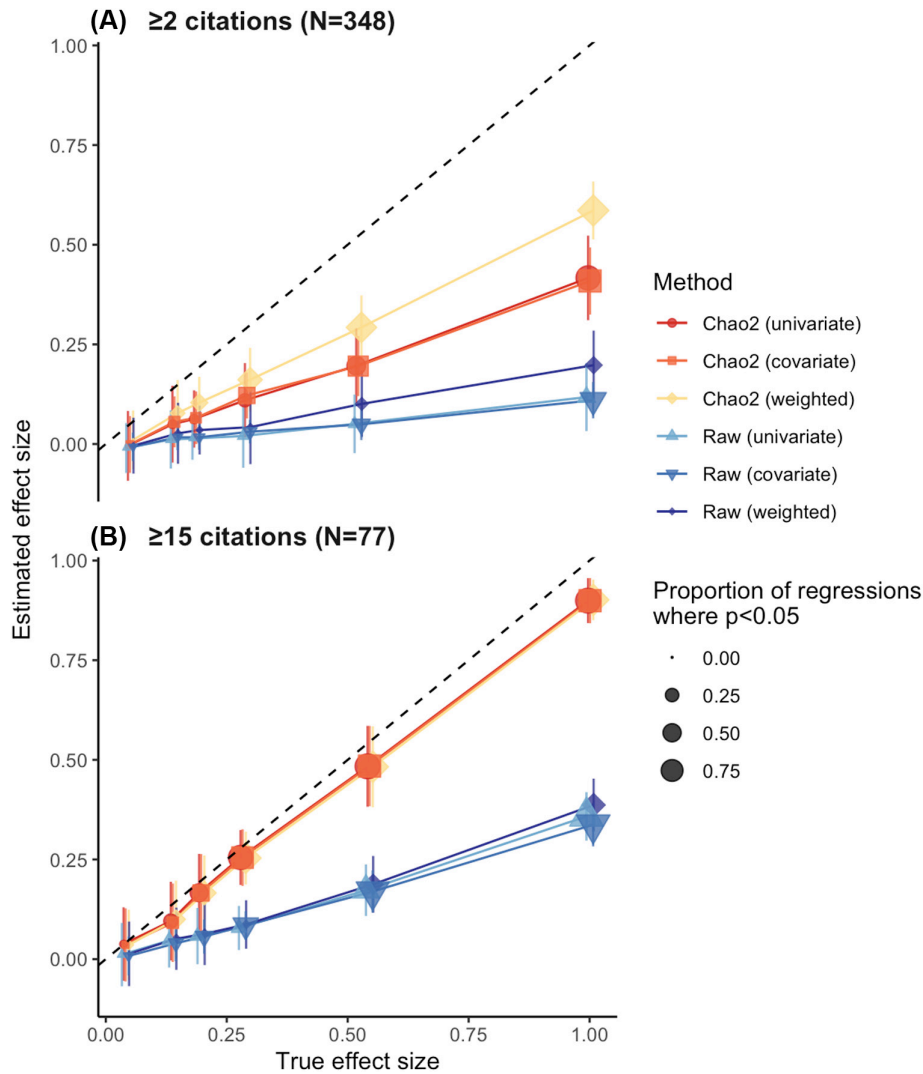
Figure 3. All modeling methods are better able to recover a true relationship between a host trait and PSR increase as the true effect size increases. The effect size is the estimated slope from linear models of (estimated or raw) PSR based on the trait and sampling intensity (except in the univariate models). The size of each point indicates the proportion of simulations where the coefficient was significant at $p < 0.05$. The dashed line shows a 1:1 line. All models of the Chao2 estimation (yellow, red and orange points) perform better than those using raw (reported) PSR (blue points), but all methods substantially underestimate effect sizes. Increasing the citation cutoff from 2 (A) to 15 (B) slightly increases effect sizes and the probability of detecting a true relationship.

for parasitological databases. Accordingly, other studies have shown that the relative performance of the Chao2 and jackknife estimators depends on the variation in true species richness (Rajakaruna et al. 2016) and/or sample size (Unterseher et al. 2008).

Even though Chao2 performed better than other estimators, estimates of PSR were generally imprecise (i.e. $R^2 < 0.1$ and only up to $R^2 < 0.50$ when assuming realistic patterns of sampling, Fig. 2). Bias in parasite sampling, in which some parasite species were more frequently re-sampled than others, was one major driver of the poor performance of the estimators. In our simulations, estimates of PSR were more accurate when sampling was unbiased in terms of the parasite species sampled (Supplementary material Appendix 1 Table A1). This

bias is not well accounted for by diversity estimators, which assume that each sample is a random subset of the individuals present, so that the likelihood of a species being sampled depends solely on its relative abundance (Chao et al. 2009). However, studies of parasites rarely aim to record all parasites on the sampled hosts (or a random selection of them). Even with increased sampling effort, there are inherent limitations in sampling total parasite diversity (e.g. a need for specialized primers for PCR, or limited sampling of different organs for highly localized parasites) (McClintock et al. 2010).

Based on these concerns, we suggest that statistical techniques are needed that explicitly account for the nature of sampling in the primary studies that comprise parasite databases. For instance, incorporating information about the
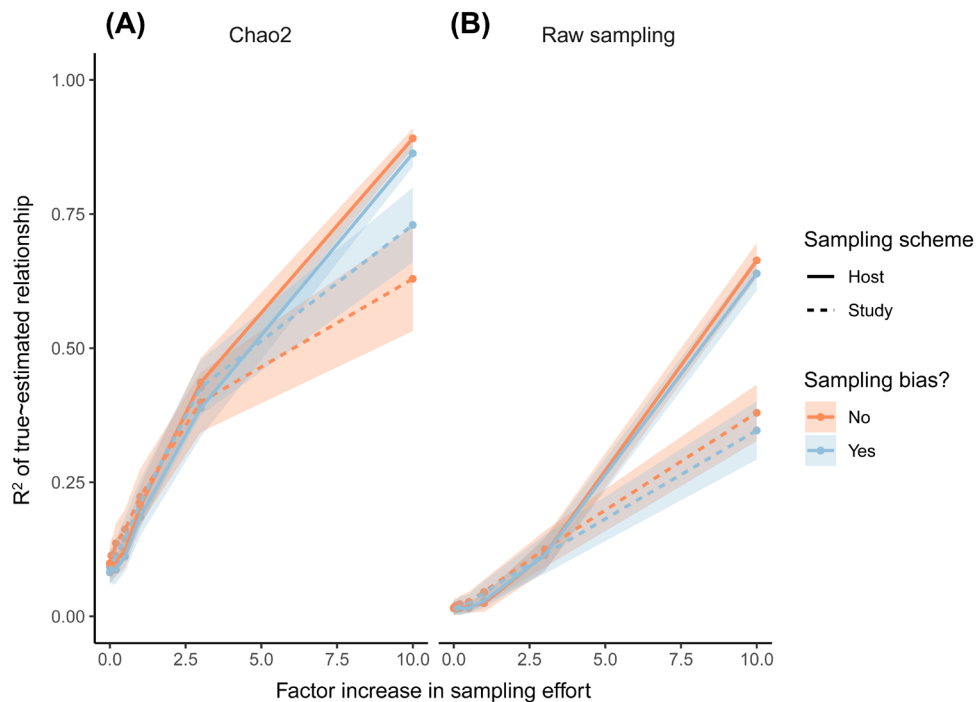
Figure 4. Performance of the Chao2 estimator (A) and raw sampling (B) improve as sampling effort increases and the sampled PSR approaches true PSR in all sampling scenarios, including increasing sampling at the level of the host species (solid lines) or study (dotted lines) and allowing bias in sampling of parasite species (blue) or not (red). Chao2 always outperforms raw sampling, even at high levels of sampling.

identities of the investigators (who may specialize in studying certain parasites), biological information on host species (e.g. detection probabilities, proximity to human populations), or parasite ecology (Cooper and Nunn 2013) could give a more mechanistic view of the processes that underlie biased sampling across host and parasite taxa (Olival et al. 2017). Statistical methods could also account for differences in the sensitivity of different detection methods; for example, detecting parasite DNA is likely to produce fewer false negatives, but more false positives, than microscopy, while serological methods detect previous infections (Jarvi et al. 2002). In addition, studies could explore other metrics of sampling bias, such as the number of host individuals sampled (Figuerola and Green 2000, Shaw et al. 2018), or bias towards studying certain parasite species. When used in combination with detection probabilities, this component of sampling effort could even more accurately inform the number of undetected parasites in a host species, so future studies with information on these variables could incorporate them into measurements of sampling effort. These statistical techniques could also be applied to other non-parasite databases where sampling bias and imperfect detection are prevalent (Leles et al. 2019).

Despite the poor performance of current diversity estimators for quantifying PSR, our analyses show that these estimators can more accurately identify drivers of host PSR in comparative studies. We found only a slightly elevated risk of detecting a significant relationship when none existed; however, estimated effect sizes were almost always far lower

than true effect sizes. Spurious relationships are associated with inflated effect sizes (Loken and Gelman 2017); thus, by systematically underestimating PSR, diversity estimators artificially shrink effect sizes and thereby reduce the risk of reporting spurious results. In other words, current methods for estimating richness may prevent us from detecting weaker relationships, but these methods are unlikely to be reporting spurious relationships. Meta-analyses and reviews are also a powerful way to detect consistent relationships across studies (Rifkin et al. 2012, Kamiya et al. 2014a, b) and could thus serve as a complementary research strategy.

Recommendations

We draw several methodological recommendations from our analyses. First, we recommend that comparative studies using parasite databases, such as the GMPD, use nonparametric diversity estimators rather than rarefaction or other methods to account for biased sampling effort. Our results lend the strongest support for the Chao2 and first- and second-order jackknife estimators, potentially in combination with a weighted regression. We also note that using the Chao2 consistently under-estimated effect sizes of the relationship between PSR and a host trait, so researchers should keep in mind the substantial bias in these estimates.

Second, because diversity estimators performed poorly at recovering the true value of PSR, we caution against using these methods to estimate true PSR, and instead suggest using them for studies comparing relative values of PSR across host

species. Estimator bias was especially high in simulations where true PSR was higher (see also Walther and Morand 1998), so we recommend using estimators more cautiously in systems known to be species-rich. If it is necessary to quantify species richness directly, we recommend using the Chao2 or jackknife estimators, ensuring that a host species has been the subject of at least 15 studies, and that these studies explored a range of parasite species.

Third, performance of diversity estimators improved when limiting the dataset to host species that have been relatively well studied. Estimates for poorly-studied hosts could be more strongly affected by random variation in the counts of singletons and doubletons than are more thoroughly-sampled hosts. In addition, limiting the study to the most-studied host species could circumvent biases in parasite sampling, for example, due to a focus on more virulent parasites (Jennelle et al. 2007). For these reasons, we suggest that it is more reliable to use biodiversity estimators with a minimum citation cutoff. However, we note that sampling effort may be systematically biased towards species with certain traits (Cooper and Nunn 2013) or certain geographic areas (Olival et al. 2017, Jorge and Poulin 2018). For instance, sampling effort can correlate with the body mass or geographic range of a host species (Nilsen and Linnell 2006). In this case, introducing a citation cutoff would introduce additional biases, for example, in limiting analyses to larger-bodied or broader-ranging species, both of which can be important predictors of PSR (reviewed in Stephens et al. 2016).

Finally, our study highlights the value of exploratory field research for improving our ability to estimate and understand biodiversity. In particular, we suggest that new, exploratory primary studies (i.e. those that look for a broad range of previously-unsampled parasites) would be particularly valuable, as would those that focus on poorly-sampled host species, for reducing bias and improving estimator performance. These recommendations echo those for free-living species, where a focus on understudied taxa and geographic areas would enhance comparative studies (Leles et al. 2019). In both cases, sampling might be more feasible if done opportunistically; for instance, blood samples collected for other purposes could be screened for parasites by collaborating scientists without requiring additional field work. Similarly, samples can be tested for more parasites by using multiple techniques on a single sample or by using metagenomics techniques (Quince et al. 2017). Adopting such approaches would also give information on absences, which can be just as important for biodiversity research and distribution modeling (Brotons et al. 2004), but are rarely documented.

Conclusions

Our simulations allowed a rigorous assessment of methods for accounting for sampling bias in global parasite databases, with applications for other compiled databases that are also irregularly sampled (Troia and McManamay 2016). This simulation approach could be applied to other sampling issues, including spatial and temporal biases in sampling (Bates et al. 2015).

Our findings suggest that the sampling biases inherent in large databases result in poor performance of diversity estimators, but that steps can be taken to overcome these biases to meaningfully compare patterns of biodiversity across taxa, space, and time.

References

Acevedo-Whitehouse, K. et al. 2003. Evidence of *Leptospira interrogans* infection in California sea lion pups from the Gulf of California. – J. Wildl. Dis. 39: 145–151.

Altizer, S. et al. 2007. Do threatened hosts have fewer parasites? A comparative study in primates. – J. Anim. Ecol. 76: 304–314.

Bates, A. E. et al. 2015. Distinguishing geographical range shifts from artefacts of detectability and sampling effort. – Divers. Distrib. 21: 13–22.

Berger, S. A. 2005. GIDEON: a comprehensive web-based resource for geographic medicine. – Int. J. Health Geogr. 4: 10.

Blakeslee, A. M. H. and Byers, J. E. 2008. Using parasites to inform ecological history: comparisons among three congeneric marine snails. – Ecology 89: 1068–1078.

Bordes, F. et al. 2007. Rodent sociality and parasite diversity. – Biol. Lett. 3: 692–694.

Brose, U. et al. 2003. Estimating species richness: sensitivity to sample coverage and insensitivity to spatial patterns. – Ecology 84: 2364–2377.

Brotons, L. et al. 2004. Presence–absence versus presence-only modelling methods for predicting bird habitat suitability. – Ecography 27: 437–448.

Burnham, K. P. and Overton, W. S. 1978. Estimation of the size of a closed population when capture probabilities vary among animals. – Biometrika 65: 625–633.

Carlson, C. J. et al. 2019. Global estimates of mammalian viral diversity accounting for host sharing. – Nat. Ecol. Evol. 3: 1070–1075.

Chao, A. 1984. Nonparametric estimation of the number of classes in a population. – Scand. J. Stat. 11: 265–270.

Chao, A. and Yang, M. C. K. 1993. Stopping rules and estimation for recapture debugging with unequal failure rates. – Biometrika 80: 193–201.

Chao, A. et al. 2009. Sufficient sampling for asymptotic minimum species richness estimators. – Ecology 90: 1125–1133.

Chao, A. et al. 2016. SpadeR: species-richness prediction and diversity estimation with R. – <https://cran.r-project.org/web/packages/SpadeR/SpadeR.pdf>.

Cleaveland, S. et al. 2001. Diseases of humans and their domestic mammals: pathogen characteristics, host range and the risk of emergence. – Phil. Trans. R. Soc. B 356: 991–999.

Cooper, N. and Nunn, C. L. 2013. Identifying future zoonotic disease threats: where are the gaps in our understanding of primate infectious diseases? – Evol. Med. Public Heal. 2013: 27–36.

Cooper, N. et al. 2012. Phylogenetic host specificity and understanding parasite sharing in primates. – Ecol. Lett. 15: 1370–1377.

Dobson, A. et al. 2008. Homage to Linnaeus: how many parasites? How many hosts? – Proc. Natl Acad. Sci. USA 105 (Suppl. 1): 11482–11489.

Dunn, R. R. et al. 2017. The global synanthrome project: a call for an exhaustive study of human associates. – Trends Parasitol. 33: 4–7.

Edie, S. M. et al. 2017. Probabilistic models of species discovery and biodiversity comparisons. – Proc. Natl Acad. Sci. USA 114: 3666–3671.

Ezenwa, V. O. et al. 2006. Host traits and parasite species richness in even and odd-toed hoofed mammals, Artiodactyla and Perissodactyla. – Oikos 115: 526–536.

Figuerola, J. and Green, A. J. 2000. Haematozoan parasites and migratory behaviour in waterfowl. – Evol. Ecol. 14: 143–153.

Gibson, D. I. et al. 2005. Host–parasite database of the natural history museum. – <www.nhm.ac.uk/research-curation/scientific-resources/taxonomy-systematics/host-parasites/>.

Gotelli, N. J. and Colwell, R. K. 2011. Estimating species richness. – In: Magurran, A. E. and McGill, B. J. (eds), Biological diversity: frontiers in measurement and assessment. Oxford Univ. Press, pp. 39–54.

Gregory, A. R. D. 1990. Parasites and host geographic range as illustrated by waterfowl. – Funct. Ecol. 4: 645–654.

Han, B. A. et al. 2015. Infectious disease transmission and behavioural allometry in wild mammals. – J. Anim. Ecol. 84: 637–646.

Huang, S. et al. 2015. Parasite diversity declines with host evolutionary distinctiveness: a global analysis of carnivores. – Evolution 69: 621–630.

Jarvi, S. I. et al. 2002. PCR diagnostics underestimate the prevalence of avian malaria (*Plasmodium relictum*) in experimentally-infected passerines. – J. Parasitol. 88: 153–158.

Jennelle, C. S. et al. 2007. State-specific detection probabilities and disease prevalence. – Ecol. Appl. 17: 154–167.

Jorge, F. and Poulin, R. 2018. Poor geographical match between the distributions of host diversity and parasite discovery effort. – Proc. R. Soc. B 285: 20180072.

Kamiya, T. et al. 2014a. What determines species richness of parasitic organisms? A meta-analysis across animal, plant and fungal hosts. – Biol. Rev. 89: 123–134.

Kamiya, T. et al. 2014b. Host diversity drives parasite diversity: meta-analytical insights into patterns and causal mechanisms. – Ecography 37: 689–697.

Leles, S. G. et al. 2019. Sampling bias misrepresents the biogeographical significance of constitutive mixotrophs across global oceans. – Global Ecol. Biogeogr. 28: 418–428.

Lobo, J. M. et al. 2018. KnowBR: an application to map the geographical variation of survey effort and identify well-surveyed areas from biodiversity databases. – Ecol. Indic. 91: 241–248.

Loken, E. and Gelman, A. 2017. Measurement error and the replication crisis. – Science 355: 584–585.

May, R. M. 1988. How many species are there on Earth? – Science 241: 1441–1449.

McClintock, B. T. et al. 2010. Seeking a second opinion: uncertainty in disease ecology. – Ecol. Lett. 13: 659–674.

Mora, C. et al. 2011. How many species are there on earth and in the ocean? – PLoS Biol. 9: e1001127.

Nilsen, E. B. and Linnell, J. D. C. 2006. Intra-specific variation and taxa-sampling affects the home range–body mass relationship. – Acta Theriol. 51: 225–232.

Nunn, C. L. and Altizer, S. 2006. Infectious diseases in primates: behavior, ecology and evolution. – Oxford Univ. Press.

Nunn, C. L. et al. 2003. Comparative tests of parasite species richness in primates. – Am. Nat. 162: 597–614.

Nunn, C. L. et al. 2005. Latitudinal gradients of parasite species richness in primates. – Divers. Distrib. 11: 249–256.

Oksanen, J. et al. 2018. vegan: community ecology package. – <https://cran.r-project.org/web/packages/vegan/index.html>.

Olival, K. J. et al. 2017. Host and viral traits predict zoonotic spillover from mammals. – Nature 546: 646–650.

Pappalardo, P. et al. 2020. Comparing methods for mapping global parasite diversity. – Global Ecol. Biogeogr. 29: 182–193.

Poulin, R. 1998. Comparison of three estimators of species richness in parasite component communities. – J. Parasitol. 84: 485–490.

Poulin, R. and Morand, S. 2000. The diversity of parasites. – Q. Rev. Biol. 75: 277–293.

Quince, C. et al. 2017. Shotgun metagenomics, from sampling to analysis. – Nat. Biotechnol. 35: 833–844.

Rajakaruna, H. et al. 2016. Optimizing performance of nonparametric species richness estimators under constrained sampling. – Ecol. Evol. 6: 7311–7322.

Rifkin, J. L. et al. 2012. Do animals living in larger groups experience greater parasitism? A meta-analysis. – Am. Nat. 180: 70–82.

Sanders, H. L. 1968. Marine benthic diversity: a comparative study. – Am. Nat. 102: 243–282.

Schaffer, G. D. et al. 1981. Helminth parasite of translocated raccoons (*Procyon lotor*) in the southeastern United States. – J. Wildl. Dis. 17: 217–227.

Shaw, A. K. et al. 2018. Metrics matter: the effect of parasite richness, intensity and prevalence on the evolution of host migration. – Proc. R. Soc. B 285: 20182147.

Shaw, D. J. et al. 1998. Patterns of macroparasite aggregation in wildlife host populations. – Parasitology 117: 597–608.

Smith, E. P. and van Belle, G. 1984. Nonparametric estimation of species richness. – Biometrics 40: 119.

Soberón, J. et al. 2007. Assessing completeness of biodiversity databases at different spatial scales. – Ecography 30: 152–160.

Stephens, P. R. et al. 2016. The macroecology of infectious diseases: a new perspective on global-scale drivers of pathogen distributions and impacts. – Ecol. Lett. 19: 1159–1171.

Stephens, P. R. et al. 2017. Global mammal parasite database ver. 2.0. – Ecology 98: 1476.

Teitelbaum, C. S. et al. 2018. Migratory behaviour predicts greater parasite diversity in ungulates. – Proc. R. Soc. B 285: 1–8.

Teitelbaum, C. S. et al. 2020. Data from: A comparison of diversity estimators applied to a database of host–parasite associations.

    – Dryad Digital Repository, <https://doi.org/10.5061/dryad.7pvmcvdqc>.

Troia, M. J. and McManamay, R. A. 2016. Filling in the GAPS: evaluating completeness and coverage of open-access biodiversity databases in the United States. – Ecol. Evol. 6: 4654–4669.

Unterseher, M. et al. 2008. Application of species richness estimators for the assessment of fungal diversity. – FEMS Microbiol. Lett. 282: 205–213.

Vavrek, M. J. 2011. fossil: palaeoecological and palaeogeographical analysis tools. – Palaeontol. Electron. 14: 16.

Walther, B. A. A. and Morand, S. 1998. Comparative performance of species richness estimation methods. – Parasitology 116: 395–405.

Walther, B. A. and Moore, J. L. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. – Ecography 28: 815–829.

Walther, B. A. et al. 1995. Sampling effort and parasite species richness. – Parasitol. Today 11: 306–310.

Walther, B. A. et al. 2005. The concepts of bias, precision and accuracy, and their use in testing the performance of species richness estimators, with a literature review of estimator performance. – Ecography 28: 815–829.

Woolhouse, M. E. J. et al. 1997. Heterogeneities in the transmission of infectious agents: implications for the design of control programs. – Proc. Natl Acad. Sci. USA 94: 338–342.

Yu, V. L. and Edberg, S. C. 2005. Global infectious diseases and epidemiology network (GIDEON): a world wide web-based program for diagnosis and informatics in infectious diseases. – Clin. Infect. Dis. 40: 123–126.