Rarefaction, Alpha Diversity, and Statistics

Amy D. Willis*

Department of Biostatistics, University of Washington, Seattle, WA, United States

Understanding the drivers of diversity is a fundamental question in ecology. Extensive literature discusses different methods for describing diversity and documenting its effects on ecosystem health and function. However, it is widely believed that diversity depends on the intensity of sampling. I discuss a statistical perspective on diversity, framing the diversity of an environment as an unknown parameter, and discussing the bias and variance of plug-in and rarefied estimates. I describe the state of the statistical literature for addressing these problems, focusing on the analysis of microbial diversity. I argue that latent variable models can address issues with variance, but bias corrections need to be utilized as well. I encourage ecologists to use estimates of diversity that account for unobserved species, and to use measurement error models to compare diversity across ecosystems.

Keywords: bioinformatics, computational biology, ecological data analysis, latent variable model, reproducibility, measurement error

1. INTRODUCTION

Alpha diversity metrics summarize the structure of an ecological community with respect to its richness (number of taxonomic groups), evenness (distribution of abundances of the groups), or both. Because many perturbations to a community affect the alpha diversity of a community, summarizing and comparing community structure via alpha diversity is a ubiquitous approach to analyzing community surveys. In microbial ecology, analyzing the alpha diversity of amplicon sequencing data is a common first approach to assessing differences between environments.

Unfortunately, determining how to meaningfully estimate and compare alpha diversity is not trivial. To illustrate, consider the following example where the alpha diversity metric of interest is strain-level richness of a microbial community (the total number of strain variants present in the environment). Suppose I conduct an experiment in which I take a sample from Environment A and count the number of different microbial taxa present in my sample. I then take a sample from Environment B, count the number of different taxa in that sample, and compare it to the number of taxa in Environment A. I am likely to observe higher numbers of different taxa in the sample with more microbial reads. The library sizes can dominate the biology in determining the result of the diversity analysis (Lande, 1996).

Rarefaction is a method that adjusts for differences in library sizes across samples to aid comparisons of alpha diversity. First proposed by Sanders (1968), rarefaction involves selecting a specified number of samples that is equal to or less than the number of samples in the smallest sample, and then randomly discarding reads from larger samples until the number of remaining samples is equal to this threshold (see Hurlbert, 1971 for a deterministic version). Based on these subsamples of equal size, diversity metrics can be calculated that can contrast ecosystems "fairly," independent of differences in sample sizes (Weiss et al., 2017).

Unfortunately, rarefaction is neither justifiable nor necessary, a view framed statistically by McMurdie and Holmes (2014) in the context of comparison of relative abundances. In this article, I discuss why unequal sample sizes appear to cause special problems in the analysis of alpha diversity. I introduce a statistical perspective on the estimation of alpha diversity, and argue that a common view of diversity indices is causing fundamental issues in comparing samples. Without advocating for any particular model of microbial sampling, I suggest a general approach to comparing microbial diversity, one which accounts for uncertainty in estimating diversity metrics. However, since estimates for alpha diversity metrics are heavily biased when taxa are unobserved, comparing alpha diversity using either raw or rarefied data should not be undertaken. I describe statistical methodology for alpha diversity analysis that adjusts for missing taxa, which should be used in place of existing common approaches to diversity analysis in ecology. While the focus of the examples is microbiome data analysis, the issues and discussion are equally applicable to macroecological data analysis. Furthermore, this discussion applies equally to diversity analyses performed at the strain, species, or other taxonomic level.

2. MEASUREMENT ERROR AND VARIANCE IN MICROBIOME STUDIES

Imagine that we had complete knowledge of every microbe in existence, including identity, abundance and location. To compare microbial diversity, we would define specific environments (e.g., the distal gut of women aged 35 living in the contiguous U.S.) and compare diversity metrics across different ecological gradients (e.g., with or without irritable bowel syndrome diagnoses). Alpha diversity could be compared exactly, because we would know entire microbial populations with perfect precision.

Unfortunately, we do not have knowledge of every microbe. We take samples from environments, and investigate the microbial community present in the sample. We use our findings about the sample to draw inferences about the environment that we are truly interested in. The samples are not of particular interest, except that they reflect the environment from which they were sampled. As we sample more and more of the environment using larger samples, we get closer to understanding the true and total microbial community of interest. This means that as we increase sampling, our calculation of any diversity metric [e.g., richness (Fisher et al., 1943), Shannon index (Shannon, 1948), and Simpson index (Simpson, 1949)] approaches the value of that diversity metric as calculated using the entire population.

Observing small samples from a large population is not an experimental set-up unique to microbial ecology: it is almost universal in statistics. The set-up where an estimate of a quantity converges to the correct value as more samples are obtained is also well understood in statistics. The unique property of microbiome experiments and alpha diversity analysis is that samples do not faithfully represent the entire microbial community under study. There is unadjusted error in using our samples as proxies for the entire community.
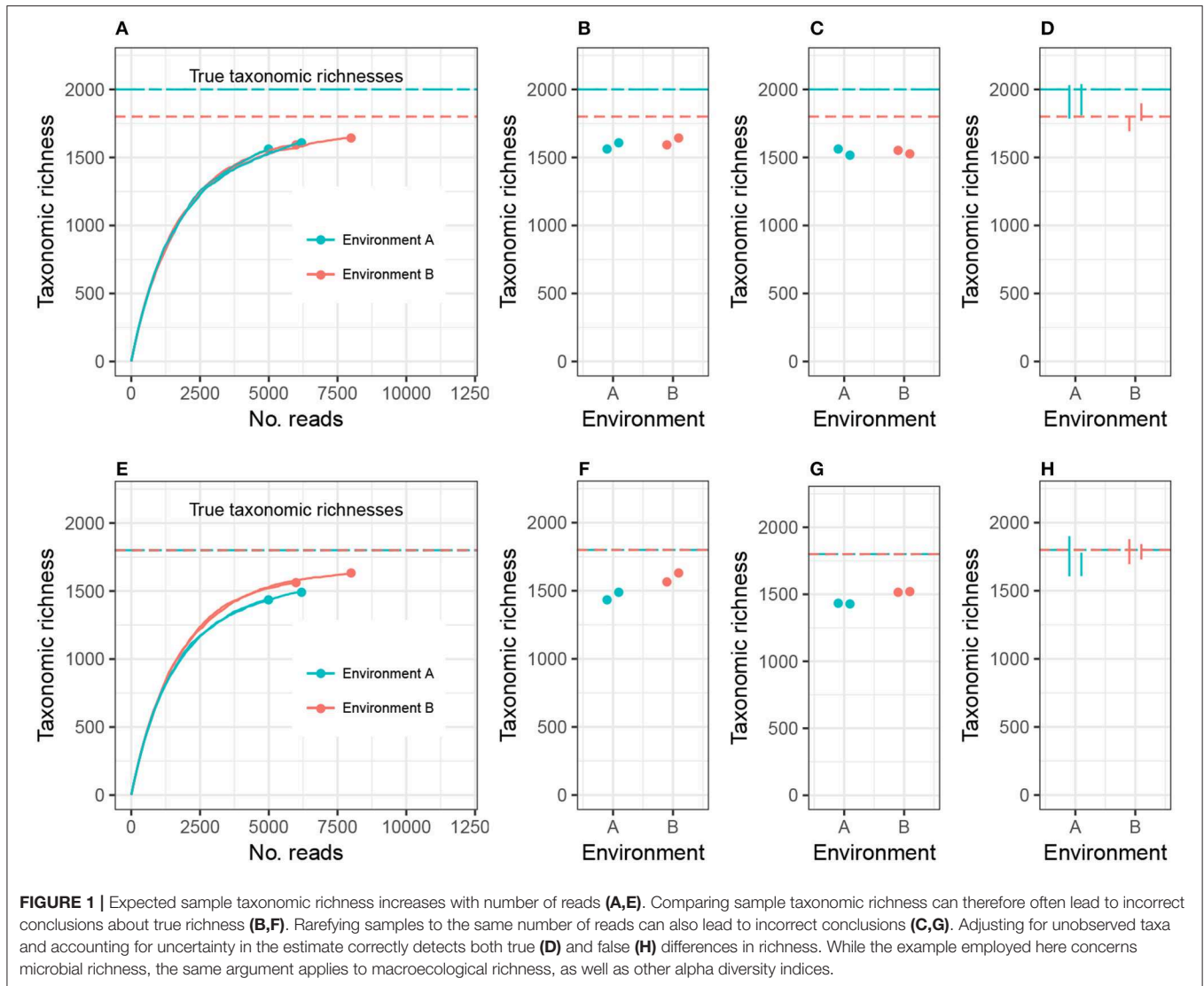
To illustrate this distinction, I contrast microbial diversity experiments with a non-microbial experiment. Suppose we are interested in modeling the CO_2 flux of soil treated with different amendments. We would measure the flux of equally sized soil sites treated with the different amendments, performing biological replicates using multiple sites for each amendment. To assess if the amendments affect the flux, we would fit a regression-type model (such as ANOVA) to flux with amendment as an explanatory variable. Implicitly, this model acknowledges that we can assess the flux with high precision; that is, the margin for error for determining flux is negligible.

Now suppose we knew that our flux-measuring machine consistently underestimated flux by exactly 5 units. We would adjust for the measurement error by adding 5 units to each measurement before comparing them. But what happens when we have random measurement error? If the measurement error on the machine was random (e.g., with 0 mean and variance of 1 unit for all amendments), this would not affect any particular amendment. However, detecting a difference between the effects of amendment on flux would be more challenging statistically: we would require more samples to detect a true difference compared to the case without measurement error. To account for the additional experimental noise, we would use a model that would account for measurement error in assessing differences between amendments. If the variance in the measurement error was 1 unit for amendment A but 5 units for amendment B, we would similarly adjust with a measurement error model.

To decide if measurement error must be accounted for when observations are made in an experiment, it is necessary to consider the effect of repeating the observational process on the same experimental unit. In the flux experiment, this would involve measuring the flux of the same soil sites again using the same experimental conditions. Without measurement error in the observations, we would consistently observe the same flux measurement, while if we had random measurement error, we would most likely observe slightly different flux measurements. Because technical replicates in microbiome experiments yield different numbers of reads, different community compositions, and different levels of alpha diversity, we have measurement error in microbial experiments. We currently do not account for measurement error in microbial diversity studies.

3. BIAS IN ESTIMATING AND COMPARING ALPHA DIVERSITY

While measurement error in microbiome studies affects all analyses of microbiome data, alpha diversity is particularly affected because commonly used estimates of alpha diversity are heavily biased compared to other estimation problems in microbial ecology (such as estimating relative abundances). Some tools to address problems with bias in alpha diversity exist in the statistical literature (Chao and Bunge, 2002; Willis and Bunge, 2015; Arbel et al., 2016; Willis and Martin, 2018). However, there are two incorrect practices surrounding alpha diversity that are preventing the uptake of statistically-motivated methodologies. The first practice is using biased estimates of alpha diversity indices. The second practice is treating alpha

FIGURE 1 | Expected sample taxonomic richness increases with number of reads **(A,E)**. Comparing sample taxonomic richness can therefore often lead to incorrect conclusions about true richness **(B,F)**. Rarefying samples to the same number of reads can also lead to incorrect conclusions **(C,G)**. Adjusting for unobserved taxa and accounting for uncertainty in the estimate correctly detects both true **(D)** and false **(H)** differences in richness. While the example employed here concerns microbial richness, the same argument applies to macroecological richness, as well as other alpha diversity indices.

diversity estimates as precisely observed quantities that do not have measurement error.

To clarify this discussion, I will focus on taxonomic richness (the simplest case), and later generalize the argument to other alpha diversity metrics. Consider the setting in **Figure 1A**, where we are investigating 2 different environments, and Environment A's richness (call it C_A) is higher than Environment B's richness (C_B). Suppose we have two biological replicates of samples from each environment: n_{A1} and n_{A2} reads from Environment A, n_{B1} and n_{B2} reads from Environment B, and $n_{A1} < n_{B1} < n_{A2} < n_{B2}$. Let c_{ij} be the observed richness of environment i on replicate j. As may commonly occur in practice, $c_{A1} < c_{A2} < c_{B1} < c_{B2}$.

There are currently two commonly used methods for comparing alpha diversity. The first method, **Figure 1B**, is to use the estimates c_{A1}, c_{A2}, c_{B1}, and c_{B2}, and perform modeling and hypothesis testing (such as ANOVA) as if both the bias and variance of these estimates were zero (see, for example, Makipaa et al., 2017). In the setting of **Figure 1A**, this leads to the erroneous conclusion that Environment A has lower richness than Environment B. The second method is to generate

a normalized, or rarefied sample by randomly discarding reads from all samples until each sample has n_{A1} reads (the number of reads in the smallest sample), **Figure 1C**. The resulting rarefied richness levels are then c_{A1}, c'_{A2}, c'_{B1}, and c'_{B2}. These estimates are then used for modeling and hypothesis testing (see, for example, Arora et al., 2017). This leads to the conclusion that Environment A and Environment B do not have significantly different richnesses, and the estimates of richness are far below the actual richnesses of each ecosystem (there is substantial negative bias in the estimates), prohibiting comparison of richness across different experiments. Furthermore, not all information collected from the samples was used in making the comparison.

Here I advocate for a third strategy: adjust the sample richness of each ecosystem by adding to it an estimate of the number of unobserved species, estimate the variance in the total richness estimate, and compare the diversities relative to these errors (**Figure 1D**). This option has the advantages of leveraging all observed reads, comparing estimates of the actual parameter of interest (taxonomic richness), and accounting for experimental noise. In the case where the environments have equal richness

(**Figures 1E–H**), this approach correctly detects equal richness, even when the abundance structures differ.

Modeling parameters observed with estimation error is not a new suggestion: this approach is from the field of statistical *meta-analysis*, where the results of multiple studies estimating the same effect size is compared (Demidenko, 2004; Willis et al., 2016; Washburne et al., 2018). In meta-analyses, larger studies need to be given more weight in determining the overall effect size, and this is incorporated into a meta-analysis via the smaller standard errors on the effect size estimates. Similarly, when comparing the response of different treatment groups in clinical trials, the number of subjects in each treatment group is accounted for in a comparison of the overall treatment effect. Adjusting for sample size when comparing different groups of observations without discarding data is widely prevalent in the sciences, and discarding data to adjust for unequal sample sizes is the exception. The strategy outlined here for modeling richness after adjusting for missing species adjusts for both bias and variance, thus accounting for library size differences and incomplete microbial surveys.

While the example discussed here is richness, this approach to estimating and comparing alpha diversity using a bias correction (incorporating unobserved taxa) and a variance adjustment (measurement error model) could apply to any alpha diversity metric. However, richness estimation has a well-studied statistical literature, and richness estimators that are adapted to microbiome data exist (see Bunge et al., 2014 for a review). The same is not true for other alpha diversity metrics. For example, the Chao-Bunge (Chao and Bunge, 2002) and breakaway (Willis and Bunge, 2015) estimators of taxonomic richness provide variance estimates, account for unobserved taxa, and are not overly sensitive to the singleton count (the number of species observed once). In contrast, the coverage adjusted entropy estimator of the Shannon index (Chao and Shen, 2003) provides variance estimates and accounts for unobserved taxa, but is extremely sensitive to the singleton count, which is often difficult to determine in microbiome studies. While alpha diversity estimation for microbiomes is an active area of research in statistics (Arbel et al., 2016; Zhang and Grabchak, 2016; Willis and Martin, 2018), there remain many features of microbial ecosystems (such as crosstalk between samples and spatial organization of microbes) that are not yet incorporated into statistical methodology for alpha diversity estimation. Despite this, alpha diversity estimates that account for unobserved taxa and provide variance estimates are vastly preferable to both plug-in and rarefied estimates, which do not account for unobserved taxa nor provide variance estimates.

4. DISCUSSION

Plug-in estimates of many alpha diversity indices (including richness and Shannon diversity) are negatively biased for the environment's alpha diversity parameter, that is, they underestimate the true alpha diversity (Lande, 1996). Attempting to address this problem using rarefaction actually induces more bias. This is sometimes justified by claiming that rarefied estimates are equally biased. However, this is not generally true, because environments can be identical with respect to one alpha diversity metric, but the different abundance structures will induce different biases when rarefied. For example, **Figure 1E** shows two environments with different abundance structures but equal richness; rarefying gives the false impression of unequal richness (see also Lande et al., 2000). In this way, both sample richness and rarefied richness are driven by artifacts of the experiment (library size), and not purely the microbial community structure. In order to draw meaningful conclusions about the entire microbial community, it is necessary to adjust for inexhaustive sampling using statistically-motivated parameter estimates for alpha diversity. In order to draw meaningful conclusions regarding comparisons of microbial communities, it is necessary to use measurement error models to adjust for the uncertainty in the estimation of alpha diversity.

It has recently been argued that studying microbial diversity without context is distracting us from gaining insight into ecological mechanisms (Shade, 2016). To this criticism, I add misapplying statistical tools is undermining many analyses of alpha diversity. I encourage microbial ecologists to use estimates of alpha diversity that account for unobserved species, and to use the variance of the estimates in measurement error models to compare diversity across ecosystems.

REFERENCES

Arbel, J., Mengersen, K., and Rousseau, J. (2016). Bayesian nonparametric dependent model for partially replicated data: the influence of fuel spills on species diversity. *Ann. Appl. Stat.* 10, 1496–1516. doi: 10.1214/16-AOAS944

Arora, T., Seyfried, F., Docherty, N. G., Tremaroli, V., le Roux, C. W., Perkins, R., et al. (2017). Diabetes-associated microbiota in fa/fa rats is modified by Roux-en-Y gastric bypass. *ISME J.* 11, 2035–2046. doi: 10.1038/ismej.2017.70

Bunge, J., Willis, A., and Walsh, F. (2014). Estimating the number of species in microbial diversity studies. *Annu. Rev.*

Stat. Appl. 1, 427–445. doi: 10.1146/annurev-statistics-022513-115654

Chao, A., and Bunge, J. (2002). Estimating the number of species in a stochastic abundance model. *Biometrics* 58, 531–539. doi: 10.1111/j.0006-341X.2002.00531.x

Chao, A., and Shen, T.-J. (2003). Nonparametric estimation of Shannon's index of diversity when there are unseen species in sample. *Environ. Ecol. Stat.* 10, 429–443. doi: 10.1023/A:1026096204727

Demidenko, E. (2004). *Mixed Models: Theory and Applications*. Hoboken, NJ: Wiley-Interscience. doi: 10.1002/0471728438

Fisher, R. A., Corbet, A. S., and Williams, C. B. (1943). The relation between the number of species and the number of individuals in a random sample of an animal population. *J. Anim. Ecol.* 12:42. doi: 10.2307/1411

Hurlbert, S. H. (1971). The nonconcept of species diversity: a critique and alternative parameters. *Ecology* 52, 577–586. doi: 10.2307/1934145

Lande, R. (1996). Statistics and partitioning of species diversity, and similarity among multiple communities. *Oikos* 76, 5–13. doi: 10.2307/3545743

Lande, R., DeVries, P. J., and Walla, T. R. (2000). When species accumulation curves intersect: implications for ranking diversity using small samples. *Oikos* 89, 601–605. doi: 10.1034/j.1600-0706.2000.890320.x

Makipaa, R., Rajala, T., Schigel, D., Rinne, K. T., Pennanen, T., Abrego, N., et al. (2017). Interactions between soil- and dead wood-inhabiting fungal communities during the decay of Norway spruce logs. *ISME J.* 11, 1964–1974. doi: 10.1038/ismej.2017.57

McMurdie, P. J., and Holmes, S. (2014). Waste not, want not: why rarefying microbiome data is inadmissible. *PLoS Comput. Biol.* 10:e1003531. doi: 10.1371/journal.pcbi.1003531

Sanders, H. L. (1968). Marine benthic diversity: a comparative study. *Am. Nat.* 102, 243–282. doi: 10.1086/282541

Shade, A. (2016). Diversity is the question, not the answer. *ISME J.* 11, 1–6. doi: 10.1038/ismej.2016.118

Shannon, C. E. (1948). A mathematical theory of communication. *Bell Syst. Tech. J.* 27, 379–423. doi: 10.1002/j.1538-7305.1948.tb01338.x

Simpson, E. H. (1949). Measurement of diversity. *Nature* 163:688. doi: 10.1038/163688a0

Washburne, A. D., Morton, J. T., Sanders, J., McDonald, D., Zhu, Q., Oliverio, A. M., et al. (2018). Methods for phylogenetic analysis of microbiome data. *Nat. Microbiol.* 3:652. doi: 10.1038/s41564-018-0156-0

Weiss, S., Xu, Z. Z., Peddada, S., Amir, A., Bittinger, K., Gonzalez, A., et al. (2017). Normalization and microbial differential abundance strategies depend upon data characteristics. *Microbiome* 5:27. doi: 10.1186/s40168-017-0237-y

Willis, A. (2017). Rarefaction, alpha diversity, and statistics. *bioRxiv* 1–8. doi: 10.1101/231878

Willis, A., and Bunge, J. (2015). Estimating diversity via frequency ratios. *Biometrics* 71, 1042–1049. doi: 10.1111/biom.12332

Willis, A. D., Bunge, J., and Whitman, T. (2016). Improved detection of changes in species richness in high-diversity microbial communities. *J. R. Stat. Soc. C Appl. Stat.* 66, 963–977. doi: 10.1111/rssc.12206

Willis, A. D., and Martin, B. D. (2018). Divnet: estimating diversity in networked communities. *bioRxiv* 1–23. doi: 10.1101/305045

Zhang, Z., and Grabchak, M. (2016). Entropic representation and estimation of diversity indices. *J. Nonparametr. Stat.* 28, 563–575. doi: 10.1080/10485252.2016.1190357