Estimating and comparing microbial diversity in the presence of sequencing errors

Chun-Huo Chiu and Anne Chao

Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan

ABSTRACT

Estimating and comparing microbial diversity are statistically challenging due to limited sampling and possible sequencing errors for low-frequency counts, producing spurious singletons. The inflated singleton count seriously affects statistical analysis and inferences about microbial diversity. Previous statistical approaches to tackle the sequencing errors generally require different parametric assumptions about the sampling model or about the functional form of frequency counts. Different parametric assumptions may lead to drastically different diversity estimates. We focus on nonparametric methods which are universally valid for all parametric assumptions and can be used to compare diversity across communities. We develop here a nonparametric estimator of the true singleton count to replace the spurious singleton count in all methods/approaches. Our estimator of the true singleton count is in terms of the frequency counts of doubletons, tripletons and quadrupletons, provided these three frequency counts are reliable. To quantify microbial alpha diversity for an individual community, we adopt the measure of Hill numbers (effective number of taxa) under a nonparametric framework. Hill numbers, parameterized by an order q that determines the measures' emphasis on rare or common species, include taxa richness ($q = 0$), Shannon diversity ($q = 1$, the exponential of Shannon entropy), and Simpson diversity ($q = 2$, the inverse of Simpson index). A diversity profile which depicts the Hill number as a function of order q conveys all information contained in a taxa abundance distribution. Based on the estimated singleton count and the original non-singleton frequency counts, two statistical approaches (non-asymptotic and asymptotic) are developed to compare microbial diversity for multiple communities. (1) A non-asymptotic approach refers to the comparison of estimated diversities of standardized samples with a common finite sample size or sample completeness. This approach aims to compare diversity estimates for equally-large or equally-complete samples; it is based on the seamless rarefaction and extrapolation sampling curves of Hill numbers, specifically for $q = 0$, 1 and 2. (2) An asymptotic approach refers to the comparison of the estimated asymptotic diversity profiles. That is, this approach compares the estimated profiles for complete samples or samples whose size tends to be sufficiently large. It is based on statistical estimation of the true Hill number of any order $q \geq 0$. In the two approaches, replacing the spurious singleton count by our estimated count, we can greatly remove the positive biases associated with diversity estimates due to spurious singletons and also make fair comparisons across microbial communities, as illustrated in our

Submitted 4 September 2015
Accepted 6 January 2016
Published 1 February 2016

Corresponding author
Anne Chao, chao@stat.nthu.edu.tw

Academic editor
Rui Feng

Additional Information and Declarations can be found on page 21

DOI **10.7717/peerj.1634**

How to cite this article Chiu and Chao (2016), Estimating and comparing microbial diversity in the presence of sequencing errors. **PeerJ 4:e1634; DOI 10.7717/peerj.1634**

simulation results and in applying our method to analyze sequencing data from viral metagenomes.

Subjects Biodiversity, Ecology, Microbiology, Statistics
Keywords Extrapolation, Hill numbers, Microbial diversity, Rarefaction, Sample coverage, Standardization, Good–Turing frequency theory, Spurious singleton, Sequencing error

INTRODUCTION

Advances in high-throughput DNA sequencing have opened a novel way to assess highly diverse microbial communities (*Sogin et al., 2006*; *Roesch et al., 2007*; *Fierer et al., 2008*; *Turnbaugh & Gordon, 2009*). However, the measurement and comparison of microbial diversity are challenging issues due to sampling limitations (*Bohannan & Hughes, 2003*; *Schloss & Handelsman, 2006*; *Schloss & Handelsman, 2008*; *Øvreås & Curtis, 2011*). These issues become more challenging when sequencing errors generate spurious low frequency counts especially singletons (*Quince et al., 2009*; *Dickie, 2010*; *Kunin et al., 2010*; *Quince et al., 2011*; *Bunge et al., 2012a*; *Bunge et al., 2012b*). In this paper, we use "species" to refer to taxa or operational taxonomic units (OTUs) under a pre-specified percentage of identity of sequences (*Schloss & Handelsman, 2005*; *Schloss & Handelsman, 2008*). We also use "individuals" to refer to sequences or any sampling unit.

In macro-ecology, Hill numbers have been increasingly used to quantify species diversity. An Ecology Forum led by *Ellison (2010)* (and papers that followed it) surprisingly achieved a consensus in the use of Hill numbers as the proper choice of diversity measure, despite intense debates existing in earlier literature regarding this issue. Hill numbers (or the effective number of species) are a mathematically unified family of diversity indices differing among themselves only by an exponent q that determines the measure's sensitivity to species relative abundances. This family includes the three most important diversity measures: species richness ($q = 0$), Shannon diversity ($q = 1$, the exponential of Shannon entropy), and Simpson diversity ($q = 2$, the inverse of Simpson index). See below for its mathematical formula and interpretation. Hill numbers were first used in ecology by *MacArthur (1965)*, developed by *Hill (1973)*, and reintroduced to ecologists by *Jost (2006)* and *Jost (2007)*. Hill numbers have been extended to incorporate evolutionary history and species traits; see *Chao, Chiu & Jost (2014)* for a recent review.

Various ecological measures have been applied to quantify the diversity of microbial communities (*Hughes et al., 2001*; *Curtis, Sloan & Scannell, 2002*). *Hill et al. (2003)* reviewed and discussed the suitability of a wide range of ecological diversity measures for use with highly diverse bacterial communities. Members of Hill numbers are also proposed as promising measures for quantifying microbial diversity. For example, *Haegeman et al. (2008)*, *Haegeman et al. (2013)* and *Haegeman et al. (2014)* recommended the use of Shannon diversity and Simpson diversity to measure and compare microbial diversity; *Doll et al. (2013)* suggested using a continuous diversity profile, a plot of Hill numbers as a continuous function of $q \geq 0$. In this paper, we adopt the general framework of Hill numbers and use continuous profiles to quantify microbial diversity. The diversity profile for $q \geq 0$ conveys all information contained in a species relative abundance distribution if community parameters (species richness and relative abundances) are known. In practice, however, community parameters are unknown and thus the true diversity must be estimated from sampling data, meaning that statistical methods are required.

In this paper, we propose two statistical approaches (non-asymptotic and asymptotic) to make fair comparisons of microbial diversity across multiple communities. A non-asymptotic approach refers to the comparison of estimated diversities of standardized samples with a common finite sample size or sample completeness (as measured by sample coverage; see below). This approach aims to compare diversity estimates for equally-large or equally-complete samples; it is based on the seamless rarefaction and extrapolation sampling curves of Hill numbers, specifically for $q = 0$, 1 and 2. Traditional sample-size-based rarefaction for species richness has been widely applied in ecology as a standardization method and also suggested by *Dickie (2010)* for molecular surveys. For species richness, *Colwell et al. (2012)* proposed an integrated rarefaction and extrapolation sampling curve for standardizing sample size; *Chao & Jost (2012)* proposed the corresponding curve for standardizing sample completeness. Hill numbers calculated from a sample, like species richness, are an increasing function of sampling effort and thus tend to increase with sample completeness. *Chao et al. (2014)* generalized previous papers (*Chao & Jost, 2012*; *Colwell et al., 2012*) on species richness to the family of Hill numbers and developed two types of standardization methods (sample-size- and sample-coverage-based). The sample-size- and sample-coverage-based integration of rarefaction and extrapolation together represent a unified non-asymptotic and non-parametric framework for estimating diversity and for making statistical inferences based on these estimates. The rarefaction and extrapolation curves for measures of small value of q (say, $0 \leq q < 2$) heavily depend on the low frequency counts, especially singletons (*Chao et al., 2014*).

An asymptotic approach refers to the comparison of the estimated asymptotic diversity profiles. That is, this approach compares the estimated profiles for samples with size tending to be sufficiently large or samples with sample completeness tending to unity. It is based on statistical estimation of the true Hill number of any order $q \geq 0$. This profile is typically generated by substituting species sample proportions into the diversity formula. However, this empirical approach generally underestimates the true profile, because samples usually miss some of the community's species due to under-sampling. Finding an analytic reduced-bias continuous diversity profile has been a long-standing challenge. *Chao & Jost (2015)* recently proposed a resolution to obtain a diversity profile estimator, which infers the asymptotic or true diversities. By applying their diversity profile estimator, the negative bias associated with the empirical diversity curve due to undetected species can be greatly reduced. The authors also used real data sets to demonstrate that the empirical and their estimated diversity profiles may give qualitatively different answers when comparing biodiversity surveys. *Chao & Jost's (2015)* diversity profile estimator for low values of q ($0 \leq q < 2$) is strongly affected by the low frequency counts. This is mainly because the observed rare species that produce low frequencies carry nearly all the information about the undetected species and play an important role in almost all statistical inferences in diversity estimation.

However, unlike macro-community ecological data, the low frequency counts, especially singletons from high-throughput DNA sequencing, are subject to various types of sequencing errors at different stages of processing (*Quince et al., 2009*; *Huse et al., 2010*; *Quince et al., 2011*). Some sequences may be misclassified as new taxa, and, accordingly, are also misclassified as singletons. Consequently, the observed singletons are greatly inflated and can comprise more than 60% of taxa in a sample (*Buee et al., 2009*). Since singletons play crucial roles in both asymptotic and non-asymptotic analyses as described above, our suggested approaches will be seriously affected if the inflated singleton count is not adjusted. A wide range of methods have been developed to reduce or correct sequencing errors (*Buee et al., 2009*; *Quince et al., 2011*) at the bioinformatics-processing stage. Without knowledge of the sources of measurement errors, statistical sampling-based methods have also recently been proposed to correct the number of spurious singletons and estimate diversity. *Bunge et al. (2012a)* and *Bunge, Willis & Walsh (2014)* proposed a parametric mixture model and a method using "left-censored" data; *Willis & Bunge (2015)* proposed an approach using the ratio of two successive frequency counts. These statistical approaches generally require different parametric assumptions about the sampling models or about the functional form of the ratio of frequency counts. Some of these parametric assumptions may not be reliably tested, and the different parametric assumptions may disallow comparisons across communities.

In this paper, we propose a novel nonparametric approach to estimate the true number of singletons in the presence of sequencing errors. Here we derive a relationship between the expected frequency of singletons and the expected frequencies of doubletons, tripletons and quadrupletons, based on a modified Good–Turing frequency formula originally developed by the founder of modern computer science Alan Turing, and his colleague *Good (1953)* and *Good (2000)*. Our estimator of singleton count is thus expressed in terms of the observed frequency counts of doubletons, tripletons and quadrupletons, provided these three frequency counts are reliable. Simulation results are reported to demonstrate an important finding about our proposed singleton count estimator. That is, when there are no sequencing errors and sample sizes are reasonably large, our estimator differs from the true singleton count only to a limited extent; when there are sequencing errors, our estimator is substantially lower than the observed singleton count. Therefore, the discrepancy between the estimated and the observed singleton counts can also be used to assess whether or not sequencing errors were present in the observed data.

Throughout the paper, "*adjusted* data/estimators" refer to those with the observed singleton count being replaced by the estimated count (the observed singleton count is discarded), whereas "*original* or *observed* data" refer to the observed data with spurious singletons possibly present. To quantify and compare microbial diversity, here we propose applying both non-asymptotic and asymptotic analyses to the adjusted data whenever the singleton count is uncertain in measurement. That is, for adjusted data, we present seamless sample-size- and coverage-based rarefaction and extrapolation sampling curves of Hill numbers (focusing on measures of $q = 0$, 1, and 2) and a continuous diversity profile estimator. Simulation results based on various taxa abundance distributions are reported to examine the performance of our method and to compare our results with those obtained from a previous ratio-based method (*Bunge, Willis & Walsh, 2014*; *Willis & Bunge, 2015*). Sequencing data from viral metagenomes (*Allen et al., 2011*; *Allen et al., 2013*) are used for illustration and comparison. The generalization of our methods to phylogenetic diversity is discussed.

METHODS

Model framework based on Hill numbers

Assume in a community that there are S species indexed by 1, 2, . . . , S, where S is an unknown parameter. Let p_i be the unknown species relative abundance of the ith species or detection probability of the ith species in any randomly observed individual, $i = 1, 2, \ldots, S$, $\sum_{i=1}^{S} p_i = 1$, and X_i be the number of individuals of the ith species detected in the sample of size n. Let f_k (abundance frequency counts), $k = 1, 2, \ldots, n$, be the number of species that are observed exactly k times or with k individuals in the sample. Here, the unobservable f_0 denotes the number of undetected species in the sample; f_1 denotes the number of singletons and f_2 denotes the number of doubletons observed in the sample.

Given a species relative abundance set $\{p_1, p_2, \ldots, p_S\}$, the Hill number of order q is defined as:

$$^{q}D = \left(\sum_{i=1}^{S} p_i^q\right)^{1/(1-q)}, \quad q \geq 0, \quad q \neq 1. \tag{1a}$$

For all $q \geq 0$, if $^{q}D = k$, then the diversity of order q of the actual community with relative abundance set $\{p_1, p_2, \ldots, p_S\}$ is the same as that of an equivalent reference community with k equally abundant species (i.e., with relative abundance set $\{1/k, 1/k, \ldots, 1/k\}$). This is why Hill numbers are referred to as the effective number of species or as species equivalents. Since the L^P norm is widely used in various disciplines, we here provide a very simple and intuitive connection between the L^P norm and Hill numbers. Note that the L^P norm for the relative abundance set of the actual community is $(\sum_{i=1}^{S} p_i^q)^{1/q}$, whereas the corresponding L^P norm for the equally abundant reference community is $(\sum_{i=1}^{k} (1/k)^q)^{1/q} = k^{(1-q)/q}$. If the two L^P norms are equal, then we have $(\sum_{i=1}^{S} p_i^q)^{1/q} = k^{(1-q)/q}$, implying $k = (\sum_{i=1}^{S} p_i^q)^{1/(1-q)}$, which is the formula of the Hill number of order q in Eq. (1a).

The parameter q determines the sensitivity of the measure to the relative abundance. When $q = 0$, the abundances of individual species do not contribute to the sum in Eq. (1a). Rather, only presences are counted, so that ^{0}D is simply species richness, which counts *species* equally without regard to their relative abundances. For $q = 1$, Eq. (1a) is undefined, but its limit as q tends to 1 is the exponential of the familiar Shannon index, referred to as Shannon diversity (*Chao et al., 2014*):

$$^{1}D = \lim_{q \to 1} {}^{q}D = \exp\left(-\sum_{i=1}^{S} p_i \log p_i\right). \tag{1b}$$

The measure for $q = 1$ counts *individuals* equally and thus counts species in proportional to their abundances; the measure ^{1}D can be interpreted as the effective number of common species in the community. The measure for $q = 2$ discounts all but the dominant species and can be interpreted as the effective number of dominant species in the community. *Hill (1973)*, *Tóthmérész (1995)*, *Gotelli & Chao (2013)*, *Doll et al. (2013)* and others suggested that biologists should use all the information contained in their data by plotting the diversity as a continuous function of $q \geq 0$. If the profiles of two communities do not cross, then one of the assemblages is unambiguously more diverse than the other. If they cross, only statements conditional on q can be made about their ranking. In most applications, the diversity profiles are plotted for all values (including non-integers) of q from 0 to $q = 3$ or 4, beyond which it generally does not change much. Thus, our diversity profile is mainly focused on the range of $0 \leq q \leq 3$.

Modified Good–Turing frequency formula

The original Good–Turing frequency formula was developed during World War II cryptographic analyses by Alan Turing and I. J. Good. Turing never published the theory but gave permission to Good to publish it. Two influential papers by *Good (1953)* and *Good & Toulmin (1956)* presented Turing's wartime statistical work on the frequency formula and related topics. In an ecological context, the Good–Turing frequency theory answers a question as follows: For those species that appeared r times, $r = 0, 1, \ldots$, in a sample of size n, how can one estimate the true mean relative abundance α_r of those species? Good and Turing focused on the case of small r, i.e., rare species (or rare code elements, in Turing's case). Mathematically, $\alpha_r = \sum_{i=1}^{S} p_i I(X_i = r)/f_r$, where $I(A)$ is the indicator function, i.e., $I(A) = 1$ if the event A occurs, and 0 otherwise. Ecologists have been using the sample fraction r/n to infer α_r, but the Good–Turing frequency formula states that α_r should be estimated by r^*/n, where $r^* = (r+1)f_{r+1}/f_r$. That is, their estimator is

$$\tilde{\alpha}_r = \frac{(r+1)}{n}\frac{f_{r+1}}{f_r} \equiv \frac{r^*}{n}, \quad r = 0,\ 1, \ldots, \tag{2a}$$

The above Good–Turing frequency formula has found a wide range of applications in biological sciences, statistics, computer sciences, information sciences, and linguistics, among others. *Good (1953)* used a Bayesian approach to theoretically justify Eq. (2a) whereas *Robbins (1968)* derived it as an empirical Bayes estimator. *Good (2000)* wrote "when preparing my 1953 article, I had forgotten Turing's somewhat informal proof in 1940 or 1941, which involved cards or urn models in some way, and I worked out a separate proof (Bayes estimator). I still don't recall Turing's proof." Nevertheless, *Good (1983*, p. 28) provided a very intuitive justification of the Good–Turing frequency formula as follows: Given an original sample of size n, consider the probability of the event that the next individual will be a species that had appeared r times in the original sample. (Mathematically, this probability is simply $\sum_{i=1}^{S} p_i I(X_i = r) = \alpha_r f_r$). If this event occurs, then the species to which the additional individual belongs must appear r+1 times in the enlarged sample of size n+1. Since the order in which individuals were sampled is assumed to be irrelevant, the total number of individuals in the enlarged sample of size n+1 for those species (that had appeared r times in the original sample) is $(r+1)f_{r+1}$. Thus, the probability of the aforementioned event in the enlarged sample of size n+1 is

$(r+1)f_{r+1}/(n+1) \approx (r+1)f_{r+1}/n$. Dividing this by the number of such species, f_r, we obtain the mean relative abundance α_r of those species, which is given in Eq. (2a). *Chiu et al. (2014)* modified the Good–Turing estimator to obtain a more accurate estimator:

$$\hat{\alpha}_r = \frac{(r+1)f_{r+1}}{(n-r)f_r + (r+1)f_{r+1}}, \quad r = 0,\ 1, \ldots. \tag{2b}$$

This modified formula will be used below in deriving our estimator of the true singleton count.

Singleton count estimation

An intuitive and basic concept in estimating the number of undetected species is that abundant species (which are certain to be detected in samples) contain almost no information about undetected species richness, whereas rare species (which are likely to be either undetected or infrequently detected) contain almost all the information about undetected species richness. Therefore, most nonparametric estimators of the number of undetected species are based on counts of detected rare species, especially the numbers of singletons and doubletons. *Chao (1984)* derived a lower bound of undetected species richness in terms of counts of singletons and doubletons; the corresponding lower bound of species richness given below is referred to as the *Chao1 estimator*: (*Colwell & Coddington, 1994*)

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

Applying a similar concept and derivation, we propose below an estimator of singleton count. Given $\{p_1, p_2, \ldots, p_S\}$ a general expectation formula for the k-th order frequency count is:

$$E(f_k) = \sum_{i=1}^{S} \binom{n}{k} p_i^k (1-p_i)^{n-k}, \quad k = 0,\ 1,\ \ldots,\ n. \tag{3}$$

Based on this formula, the Cauchy-Schwarz inequality

$$\left(\sum_{i=1}^{S} p_i(1-p_i)^{n-1}\right)\left(\sum_{i=1}^{S} p_i^3(1-p_i)^{n-3}\right) \geq \left(\sum_{i=1}^{S} p_i^2(1-p_i)^{n-2}\right)^2$$

leads to

$$\frac{E(f_1)}{n} \times \frac{6E(f_3)}{n(n-1)(n-2)} \geq \left(\frac{2E(f_2)}{n(n-1)}\right)^2,$$

which implies

$$E(f_1) \geq \frac{2(n-2)[E(f_2)]^2}{3(n-1)E(f_3)}. \tag{4a}$$

Replacing the expectation terms by observed data, we obtain a preliminary lower bound for the true singleton frequency count:

$$\tilde{f}_1 = \frac{2(n-2)(f_2)^2}{3(n-1)f_3}. \tag{4b}$$

To obtain a more accurate estimator, we evaluate the magnitude of the bias of the preliminary lower bound in Eq. (4b) as

$$\left|bias\ (\tilde{f}_1)\right| \approx E(f_1) - \frac{2(n-2)[E(f_2)]^2}{3(n-1)E(f_3)}.$$

Using the definition of α_r in the Good–Turing frequency formula, we obtain the following two approximation formulas:

$$\frac{E(f_1)}{n} = \sum_{i=1}^{S} \frac{1-p_i}{p_i}\binom{n}{2}^{-1} E[I(X_i = 2)] \approx \frac{1-\alpha_2}{\alpha_2}\binom{n}{2}^{-1} E(f_2),$$

$$\frac{2E(f_2)}{n(n-1)} = \sum_{i=1}^{S} \frac{1-p_i}{p_i}\binom{n}{3}^{-1} E[I(X_i = 3)] \approx \frac{1-\alpha_3}{\alpha_3}\binom{n}{3}^{-1} E(f_3).$$

Substituting the above two approximations into the bias formula, we obtain the magnitude of bias:

$$\left|bias\ (\tilde{f}_1)\right| \approx \frac{2}{n-1}\left(\frac{1-\alpha_2}{\alpha_2} - \frac{1-\alpha_3}{\alpha_3}\right)E(f_2).$$

The right hand side of the above formula will be positive for reasonably large sample sizes, because species that are observed three times in a sample should have a larger mean abundance than that of doubletons (i.e., α_3 is larger than α_2). Applying the modified Good–Turing estimates in Eq. (2b) for α_3 and α_2, we then obtain an estimator of the true number of singletons in terms of (f_2, f_3, f_4) for large sample size n:

$$\hat{f}_1 = \frac{2f_2^2}{3f_3} + 2f_2\left(\frac{f_2}{3f_3} - \frac{f_3}{4f_4}\right). \tag{5}$$

When there are spurious singletons, we can adjust the Chao1 estimator (*Chao, 1984*) by replacing the observed singleton count f_1 with the estimated singleton count $\hat{f}_1$. Then we have the Chao1 estimator of species richness based on the adjusted data if $f_2 > 0$:

$$\hat{S}_{adjChao1} = S_{obs} - f_1 + \hat{f}_1 + \frac{(n-1)}{n}\frac{\hat{f}_1^2}{2f_2}, \tag{6a}$$

where S_{obs} denotes the number of species in the original data. When $f_2 = 0$, a bias-corrected estimator is suggested:

$$\hat{S}^*_{adjChao1} = S_{obs} - f_1 + \hat{f}_1 + \frac{\hat{f}_1(\hat{f}_1 - 1)}{2(f_2 + 1)}. \tag{6b}$$

The variance of the adjusted Chao1 estimator and the corresponding 95% confidence intervals via a log normal transformation can be obtained using similar derivations as those for the classic Chao1 estimator (*Chao, 1987*).

Non-asymptotic approach: rarefaction and extrapolation based on adjusted data

It is well known that species richness based on sampling data is highly dependent on sample size and sample completeness (*Colwell & Coddington, 1994*). *Chao et al. (2014)* showed that empirical Shannon diversity is moderately dependent and that Simpson diversity is weakly dependent on sample size and inventory completeness. They proposed two standardization methods for Hill numbers to compare non-asymptotic diversities across multiple assemblages as described below. For each type of standardization, we here mainly focus on the three measures of $q = 0$, 1 and 2 based on the adjusted data.

1. Sample-size-based rarefaction and extrapolation up to a maximum size. For each diversity measure, we standardize all samples by estimating diversity for a standard sample size, which can be smaller than an observed sample (traditional rarefaction) or larger than an observed sample (extrapolation). Then we construct for each sample an integrated rarefaction and extrapolation sampling curve as a function of sample size. For species richness, the size can be extrapolated at most to double or triple the minimum observed sample size. For Shannon diversity and Simpson diversity, if data are not too sparse, the extrapolation can be reliably extended to infinity to attain the estimated asymptote given below in Eq. (7).
2. Coverage-based rarefaction and extrapolation up to a maximum coverage. *Chao & Jost (2012)* proposed standardizing samples by matching their sample completeness, which is measured by *sample coverage*, an objective measure of sample completeness due to Turing and *Good (1953)* and *Good (2000)*. The sample coverage of a given sample is defined as the fraction of the individuals in an assemblage that belong to the species observed in the sample. Contrary to intuition, sample coverage for the observed sample, rarified samples, and extrapolated samples can be accurately estimated by the observed data themselves. The coverage-based rarefaction and extrapolation curve plots the diversity estimates as a function of sample coverage up to a maximum coverage. For species richness, the maximum coverage is selected as the coverage of the maximum size used in the sample-size-based sampling curve. For Shannon diversity and Simpson diversity, if data are not sparse, the extrapolation can often be extended to the coverage of unity to attain the estimated asymptote given below in Eq. (7).

Chao et al. (2014) introduced a bootstrap method to construct 95% confidence intervals associated with each estimated diversity measure. Generally, for any fixed sample size or any degree of completeness in the comparison, if the 95% confidence intervals do not overlap, then significant differences at a level of 5% among the expected diversities (whether interpolated or extrapolated) are guaranteed. However, overlapped intervals do not guarantee non-significance (*Colwell et al., 2012*); in this case, data are inconclusive.

The sample-size-based approach plots the estimated diversity as a function of sample size, whereas the corresponding coverage-based approach plots the same diversity with respect to sample coverage. Therefore, the two types of sampling curves can be bridged by a *sample completeness curve*, which shows how the sample coverage varies with sample size and also provides an estimate of the sample size needed to achieve a fixed degree of completeness. This curve and all the rarefaction and extrapolation estimators along with their confidence intervals can be obtained using the R package "iNEXT" which can be also downloaded from Anne Chao's website at http://chao.stat.nthu.edu.tw/software-download/.

Asymptotic approach: diversity profile estimation based on adjusted data

The *Chao & Jost (2015)* diversity profile estimator of ${}^qD$ (Eq. 1a) based on the adjusted singleton count $\hat{f}_1$ and the original non-singleton frequency counts can be expressed as

$${}^q\hat{D}_{adj} = \left(\sum_{k=0}^{n-1} \binom{q-1}{k} (-1)^k \hat{\Delta}(k) + \frac{\hat{f}_1}{n}(1-A)^{-n+1} \left[A^{q-1} - \sum_{r=0}^{n-1} \binom{q-1}{r} (A-1)^r \right] \right)^{1/(1-q)}, \; q \geq 0, \tag{7}$$

where $\hat{\Delta}(0) = 1$,

$$\hat{\Delta}(k) = \sum_{1 \leq X_i \leq n-k} \frac{\binom{n-k-1}{X_i - 1}}{\binom{n}{X_i}} = \sum_{1 \leq j \leq n-k} \frac{\binom{n-k-1}{j-1}}{\binom{n}{j}} f_j, \; k = 1, 2, \ldots, n-1,$$

and

$$A = \begin{cases} 2f_2/[(n-1)\hat{f}_1 + 2f_2], & \text{if } f_2 > 0; \\ 2/[(n-1)(\hat{f}_1 - 1) + 2], & \text{if } f_2 = 0, \hat{f}_1 \neq 0; \\ 1, & \text{if } f_2 = \hat{f}_1 = 0. \end{cases}$$

The diversity estimator of order q in each profile represents the asymptote in the rarefaction and extrapolation curves described above. To compute the profile estimator in Eq. (7) and the corresponding 95% bootstrap confidence interval, we provide R code (Supplemental Text S1) which is a modified version from the script provided in *Chao & Jost (2015)*. We consider the three special cases of $q = 0$, 1 and 2 below.

For $q = 0$, the estimator in Eq. (7) reduces to the adjusted Chao1 estimator given in Eq. (6a). Thus, it is generally a minimum number of species. For $q = 1$, the estimation of the Shannon diversity from incomplete samples is surprisingly nontrivial and has been extensively discussed in many research fields; see *Chao, Wang & Jost (2013)* for a review and a low-bias estimator. The estimator (7) for $q = 1$ reduces to their Shannon diversity estimator (given below), which can be compared across communities.

$$ {}^{1}\hat{D}_{adj} = \exp\left(\sum_{1 \le X_i \le n-1} \frac{X_i}{n} \left(\sum_{k=X_i}^{n-1} \frac{1}{k} \right) + \frac{\hat{f}_1}{n} (1-A)^{-n+1} \left[-\log A - \sum_{r=1}^{n-1} \frac{(1-A)^r}{r} \right] \right). $$

This estimator greatly reduces the negative bias associated with the empirical Shannon diversity. For $q = 2$, the Simpson diversity only counts dominant ones, and dominant species always appear in samples and undetected classes are discounted. Thus the Simpson diversity can often be accurately measured and compared across multiple communities. The estimator (7) for $q = 2$ becomes the nearly unbiased estimator of Simpson diversity (*Gotelli & Chao, 2013*):

$$ {}^{2}\hat{D}_{adj} = \left(\sum_{X_i \ge 2} \frac{X_i(X_i - 1)}{n(n-1)} \right)^{-1}. $$

Notice that singleton count is not involved in the above formula, but the sample size n is affected by the adjusted singleton count. Consequently, the effect of spurious singleton count is much less pronounced than that for measures of $q = 0$ and 1.

SIMULATION RESULTS

Since both non-asymptotic and asymptotic analyses depend on the quality of the estimated singleton count, it is essential to investigate the performance of the proposed estimator in Eq. (5). We conducted a simulation by generating data from six species abundance distributions with various degrees of heterogeneity in species relative abundances (details are provided in Supplemental Text S2). In each model, we fixed the number of species at $S = 2{,}000$ to mimic microbial communities. Then for each given model, we considered a range of sample sizes ($n = 2{,}000$ to 10,000 in an increment of 2,000). The degree of heterogeneity in species relative abundances is quantified by the CV (which is the ratio of the standard deviation over the mean) of species relative abundances. When all species relative abundances are equal, CV = 0. A larger value of CV indicates a higher degree of heterogeneity among species abundances.

For each combination of abundance model and sample size, we generated two types of data: (i) data without sequencing errors, and (ii) spurious data with a sequencing error rate of 10%, i.e., there was a 10% chance that a sampled individual was misclassified to a new species and thus became a spurious singleton. In Fig. 1, we show the plots of the average values (over 1,000 simulation trials) of four singleton counts as a function of sample size that was used in data generation. The four singleton counts include the true singleton count generated from the data without sequencing error, the spurious singleton count generated from the data with sequencing error, the adjusted singleton count based on Eq. (5), and the count obtained from the ratio-based method of *Bunge, Willis & Walsh (2014)* and *Willis & Bunge (2015)* through the R package "breakaway," available from CRAN (Comprehensive R Archive Network). In Fig. 2, the corresponding root mean squared errors (RMSEs) for the ratio-based and the proposed methods are shown. The patterns revealed by these plots are summarized below.

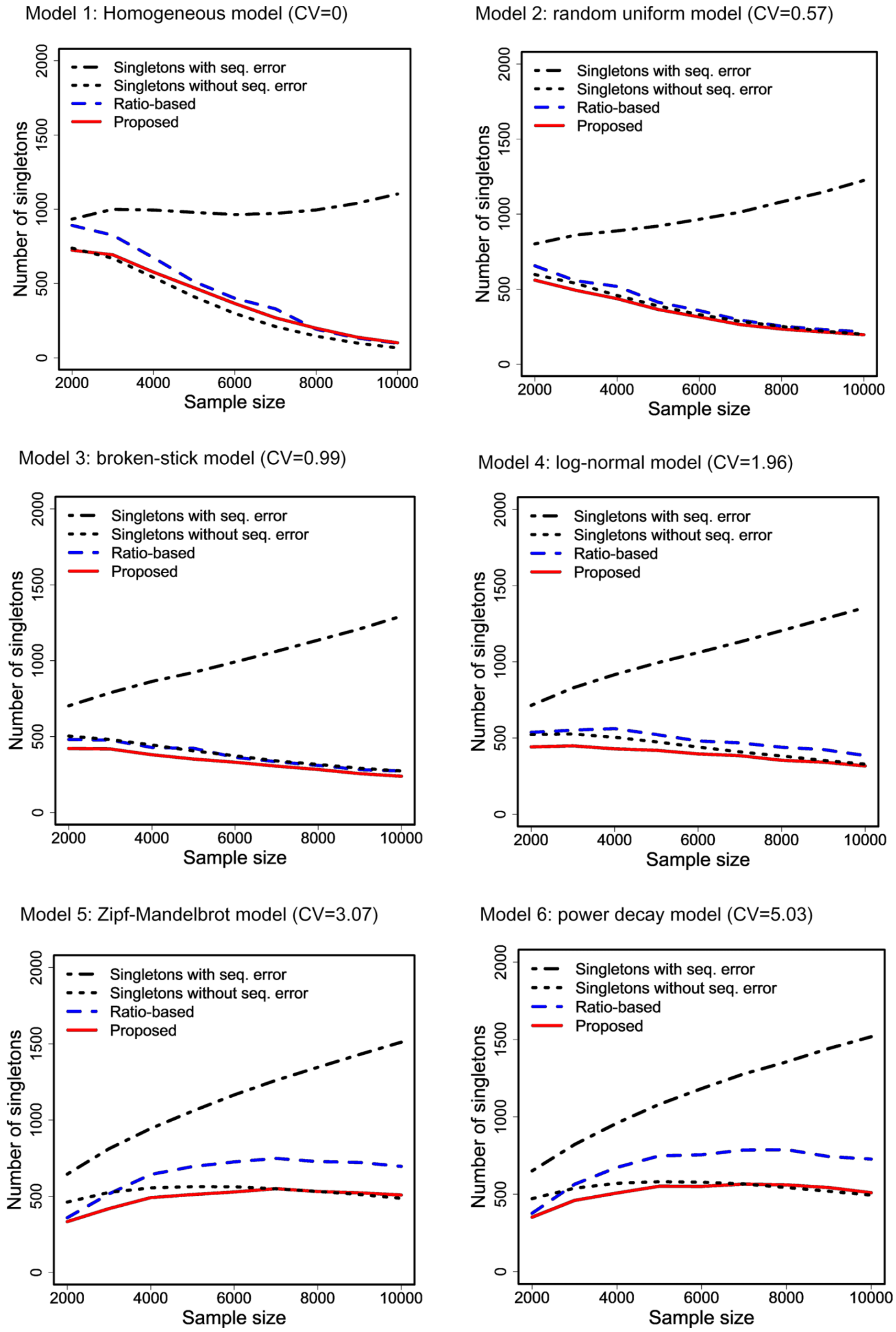

Figure 1 **Comparison of the average values of four singleton counts as a function of sample size that was used in data generation.** The four singleton counts include the true singleton count generated from the data without sequencing error, the spurious singleton count generated from the data with sequencing error, the adjusted singleton count based on Eq. (5), and the count obtained from the ratio-based method of *Bunge, Willis & Walsh (2014)* and *Willis & Bunge (2015)* through the R package "breakaway," available from CRAN. All values are averaged over 1,000 simulation trials under six species abundance models with various degrees of heterogeneity of the species abundances, as reflected by the CV value (the ratio of the standard deviation over the mean); see Supplemental Text S2 for details.

Figure 1 reveals that the number of singletons for the data without sequencing error (dotted curve in each panel) generally declines with sample size when sample size becomes sufficiently large, whereas the number of singletons for data with sequencing error (dash-dotted curve in each panel) always increases with sample size, revealing a drastically different pattern; see *Dickie (2010)* for a similar finding. This pattern can be used to detect whether sequencing error exists in the original data when an empirical accumulation curve for the singleton count can be recorded in the data-collecting procedures.

Simulation results also show that when heterogeneity is low as reflected by low CV values (Model 1 to Model 4) the ratio-based method (dashed curve) and our proposed method (solid curve) yield similar singleton counts that are close to the true data (i.e., data without sequencing error, dotted curve). The RMSEs of the two methods are thus generally comparable (Fig. 2). However, in the highly heterogeneous cases as reflected by relatively high CV values (Model 5 and Model 6), the ratio-method produces much higher singleton counts compared to the true data and thus much larger root mean squared errors than the proposed method, as shown in Fig. 2. In these high-CV cases, our estimator of singleton count still closely matches the true number of singletons, although it exhibits negative bias when sample size is relatively small especially when species abundances are highly heterogeneous.

These simulation results thus imply (i) when there are no sequencing errors (so that the dotted curves represent the singleton counts for data), our estimator differs only to a limited extent from the true data, yielding almost the same diversity inference; (ii) when there are sequencing errors (so that the dash-dotted curves represent the spurious singleton counts for data), our estimator can greatly reduce the raw singleton count and make proper corrections. Therefore, the discrepancy between our proposed estimator of singleton count and the singleton count from the observed data can be used to assess whether sequencing errors were present in data processing. Moreover, this implies that whenever the singletons are uncertain or in doubt, it is worth applying our proposed estimator of singleton count. More simulation results on the effect of spurious singletons on the estimation of asymptotic diversities are provided in Supplemental Text S2; see Discussion.

APPLICATION RESULTS

A number of data sets on frequency counts of contig (contiguous groups of sequences) spectra of viral phage metagenomes from similar or different environments were analyzed in *Allen et al. (2013)*. We select two samples with different environments to illustrate the use of our methods: one sample includes the pooled contig spectra from seven non-medicated swine feces, and the other sample includes the pooled contig spectra from four reclaimed fresh water samples. For simplicity, these two samples/viromes are respectively referred to as “swine feces” sample/virome and “reclaimed water” sample/virome in the following analysis. The frequency counts for the two samples originally provided in the additional file of *Allen et al. (2013)* are reproduced in Table 1. The empirical diversities and asymptotic analyses are shown in Table 2.

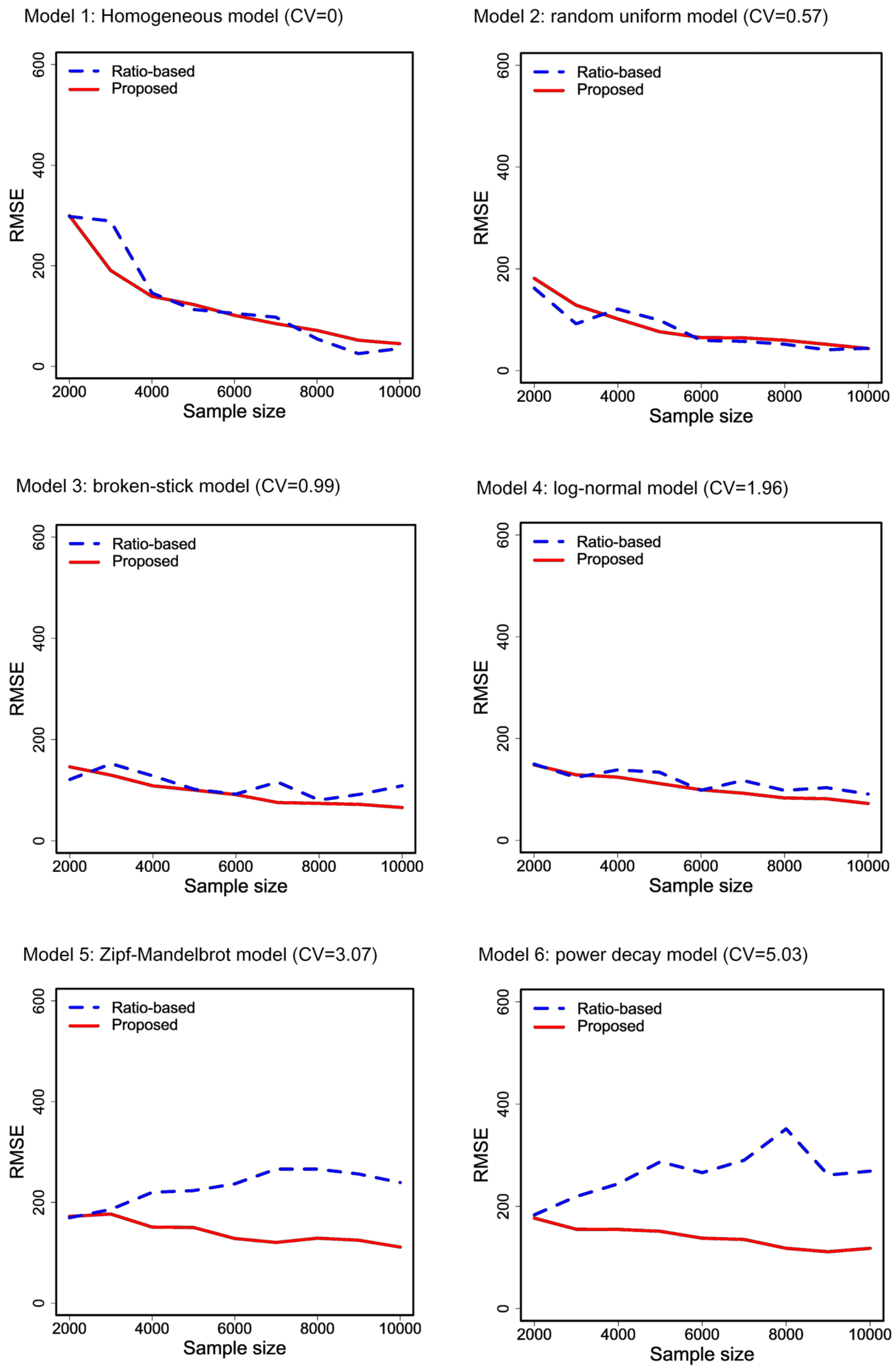

Figure 2 Comparison of the average root mean squared error (RMSE) for two singleton counts (the proposed and ratio-based estimators) as a function of the sample size that was used in data generation. The proposed method is based on Eq. (5), and the results for the ratio method (*Bunge, Willis & Walsh, 2014*; *Willis & Bunge, 2015*) were computed using the R package "breakaway." All values are averaged over 1,000 simulation trials under six species abundance models with various degrees of heterogeneity of the species abundances, as reflected by the CV value (the ratio of the standard deviation over the mean); see Supplemental Text S2 for details.

Table 1 **Frequency counts on contig spectra of phage metagenomic data (*Allen et al., 2011*; *Allen et al., 2013*).**

Sample	Original n	Adj. n	f_1	$\hat{f}_1$	f_2	f_3	f_4	f_5	f_6	f_7	f_8	f_9	f_{10}	f_{11}	f_{12}	f_{13}
Swine feces	9,988	4,794	8,025	2,831	605	129	41	16	8	4	2	1	1	1	0	0
Reclaimed water	9,973	4,092	7,986	2,105	518	129	50	24	12	7	5	3	2	1	1	1

Notes:
Swine feces sample, pooled data from seven swine non-medicated feces; Reclaimed water sample, pooled data from four reclaimed water samples; f_k, number of taxa with k sequences in the original data; $\hat{f}_1$, estimated number of singletons based on Eq. (5); Adj. n, sample size based on the adjusted data (i.e., the original data with the observed singleton count being replaced by the estimated value).

Table 2 **Comparison of empirical diversities and estimated diversities (with SE) for the original data, the adjusted data, and two previous methods, based on the phage metagenomics data (Table 1).** Previous methods include species richness estimates obtained from CatchAll software (*Allen et al., 2013*) and from a ratio-based method (*Willis & Bunge, 2015*). The adjusted data are the original data with the observed singleton count being replaced by the estimated value given in Table 1.

	Original data		Adjusted data		Previous methods	
Diversity	Empirical diversity	Estimated diversity (SE)	Empirical diversity	Estimated diversity (SE)	CatchAll (SE)	Ratio-based method (SE)
Swine feces sample						
Species richness ($q = 0$)	8,833	62,057 (1,814)	3,639	10,261 (376)	1,990 (206)	846,113 (249,481)
Shannon diversity ($q = 1$)	8,289	53,835 (1,365)	3,250	9,081 (203)		
Simpson diversity ($q = 2$)	7,348	27,801 (867)	2,742	6,404 (180)		
Reclaimed water sample						
Species richness ($q = 0$)	8,739	70,299 (1,973)	2,858	7,134 (273)	1,428 (140)	53,029 (257,637)
Shannon diversity ($q = 1$)	8,066	56,853 (1,451)	2,440	5,849 (130)		
Simpson diversity ($q = 2$)	6,817	21,535 (870)	1,922	3,625 (116)		

In the swine feces original data, there were 8,833 taxa among 9,988 individuals (sequences); the number of singletons was $f_1 = 8{,}025$, and the number of doubletons was $f_2 = 605$. In the reclaimed water data, there were 8,739 taxa among 9,973 individuals, and the first two frequency counts are $f_1 = 7{,}986$ and $f_2 = 518$. In these two original samples, most of the frequencies are concentrated on singletons. Consequently, based on the original data, the Chao1 lower bounds, 62,057 and 70,299 respectively for swine feces and reclaimed water viromes, are greatly inflated due to the presence of spurious singletons. Using Eq. (5), we obtain an estimated singleton count of 2,831 for the swine feces sample, and 2,105 for the reclaimed water sample (Table 1). For each sample, the estimated singleton count is substantially less than the observed singleton count, indicating that sequencing errors were present. The empirical and estimated diversities for the original and adjusted data are shown in Table 2. We also compare in Table 2 our estimates with those based on a ratio-based method (*Bunge, Willis & Walsh, 2014*; *Willis & Bunge, 2015*), and with those proposed in *Allen et al. (2013)* based on the CatchAll software.

From Table 2, as expected, the estimated diversity (species richness, Shannon diversity and Simpson diversity) based on the adjusted data for each sample is much lower than that based on the original data. For species richness, the CatchAll software yields excessively low estimates, even lower than the observed richness of the adjusted data. The ratio-based method, however, yields extremely large estimates for the number of species. In our simulations on species richness as described in Supplemental Text S2, we show that the ratio-based method might severely overestimate the true species richness when the heterogeneity among species abundance is relatively high. The empirical CV values for the swine feces and reclaimed water samples for adjusted date are respectively 0.62 and 0.79. As there are many undetected rare species, the true CV should be much higher than the empirical CV, leading to extremely large species richness estimates for the ratio-based method. All the following analyses are based on our adjusted data, unless otherwise stated.

Before we present the non-asymptotic analyses, we plot in Fig. 3 the sample completeness curve as a function of sample size. The sample completeness of the adjusted swine feces sample is 41%, which is lower than that for the adjusted reclaimed water sample, 48.6%. When the sample size is extrapolated to a size of 10,000 (approximately double the adjusted sample size for swine feces), the coverage of the swine feces sample is increased from 41.0% to 62.9%, whereas the coverage of the reclaimed water sample is increased from 48.6% to 74.7%. For any standardized sample size, Fig. 3 shows that the sample completeness of the swine feces sample is lower than that for the reclaimed water sample of the same size.

For non-asymptotic analysis, we present in Fig. 4 the sample-size- and coverage-based rarefaction and extrapolation curves along with 95% confidence intervals for three measures: $q = 0$, 1 and 2. The sample-size-based sampling curve is extrapolated up to a maximum size of 10,000, whereas the coverage-based sampling curve is extended up to the coverage of the size 10,000, i.e., the maximum coverage is up to 62.9% for the swine feces sample and 74.7% for the reclaimed water sample.

All plots in Fig. 4 exhibit a consistent pattern, with the diversity curve for the swine feces samples lying above the curve of the reclaimed water sample. In all plots, the 95% confidence intervals for the two samples in any rarefaction/extrapolation curve are disjoint, implying a significant difference. As stated earlier, the extrapolation for Shannon and Simpson diversity, unlike that of species richness, can often be reliably extended to infinity size or complete coverage to reach the asymptotic diversity estimate. Therefore, for Shannon diversity (common taxa richness) and Simpson diversity (dominant taxa richness), the data indicate that the swine feces virome is significantly more diverse than the reclaimed water virome. This is valid not only for the standardized sample size and sample coverage values plotted in Fig. 4, but also for entire viromes. (This is also supported by the asymptotic analysis below). For species richness, the data support this conclusion up to a standardized 62.9% fraction of each virome (Fig. 4B). Beyond that, the data do not provide sufficient information for comparison. This is because the asymptotic species richness estimator is only a lower bound (as opposed to point estimates for the other two asymptotic diversities).

For the asymptotic analysis, we plot the empirical and estimated asymptotic diversity profiles along with 95% confidence intervals in Fig. 5 when q is between 0 and 3.

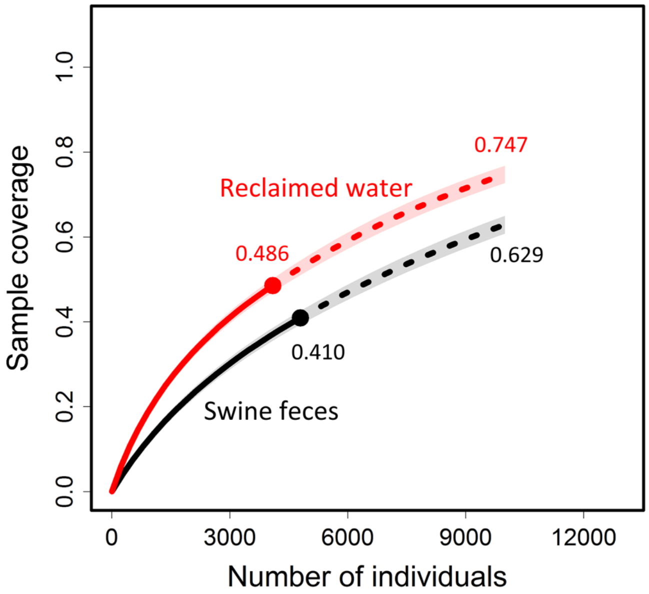

Figure 3 The sample completeness curve based on the adjusted data. Plots of sample coverage for rarefied samples (solid line) and extrapolated samples (dashed line) as a function of the sample size based on the sample frequency counts of contig spectra from seven swine fecal viromes and the sample from four reclaimed fresh water viromes (*Allen et al., 2013*). Data are given in Table 1. The original singleton count is replaced by the estimated count given in Table 1. The adjusted samples are denoted by solid dots. The 95% confidence intervals (shaded areas) were obtained by a bootstrap method based on 200 replications. Each of the two curves was extrapolated up to 10,000, approximately double the adjusted size of the swine feces sample. The numbers are the sample coverage estimates for the adjusted sample and for the sample of size 10,000.

(The empirical and estimated asymptotes of diversities for the special cases of $q = 0$, 1 and 2 are shown in Table 2, and the asymptotic diversity estimates are also shown next to an arrow at the right-hand end of each rarefaction/extrapolation plot in Fig. 5). The empirical diversities (Table 2 and Fig. 5) imply that the two viromes have limited difference in each of the three measures. In contrast, the plots in Fig. 5 reveal that for the asymptotic Shannon diversity, the swine feces virome is substantially more diverse than the reclaimed water virome. A similar conclusion is also valid for the Simpson diversity, confirming our earlier statement in the preceding paragraph.

Table 2 and Fig. 5 show that the adjusted Chao1 estimator in Eq. (6a) gives an estimate of 10,261 taxa for swine feces and 7,134 taxa for reclaimed water virome. Each is five times that obtained from CatchAll (*Allen et al., 2013*). Since the Chao1 estimate represents only minimum richness, it cannot be used to rank the taxa richness of the two entire viromes. Nevertheless, taxa richness can be compared through the coverage-based non-asymptotic approach, as discussed earlier; see Discussion. By contrast, for diversity of order $q \geq 1$, we can compare not only the estimated diversities for standardized sample size/completeness but also the estimated asymptotic diversities across communities. In Supplemental Table S1, we also give all the estimated asymptotes of diversities for other data sets provided in *Allen et al. (2013)*.

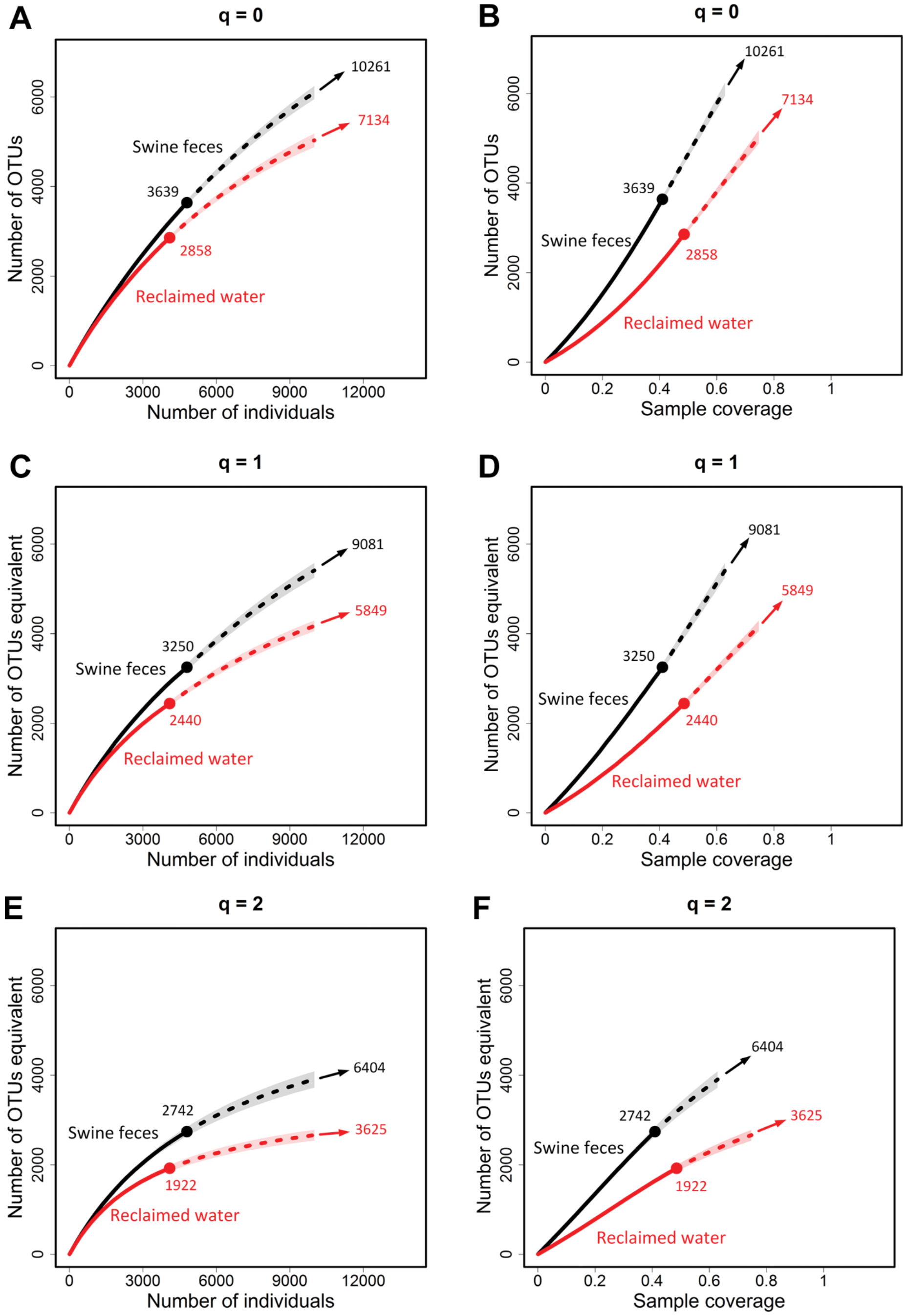

Figure 4 Non-asymptotic analysis: the rarefaction and extrapolation sampling curves based on the adjusted data. Comparison of sample-size-based (A, C, E) and sample-coverage-based (B, D, F) rarefaction and extrapolation for species richness (A, B), Shannon diversity (C, D) and Simpson diversity (E, F) based on the sample frequency counts of contig spectra from seven swine fecal viromes and the sample from four reclaimed fresh water viromes (*Allen et al., 2013*). Data are given in Table 1. The original singleton count is replaced by the estimated count given in Table 1. The adjusted samples are denoted by solid dots. Rarefied segments are denoted by solid curves and extrapolated segments are denoted by broken curves. Extrapolation is extended up to a maximum size of 10,000. Sample-coverage-based extrapolation is extended to the coverage value of the corresponding maximum sample size (i.e., 62.9% for swine feces viromes and 74.7% for reclaimed water viromes; see Fig. 3). The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. The estimated asymptotic diversity for each curve is shown next to the arrow at the right-hand end of each curve.

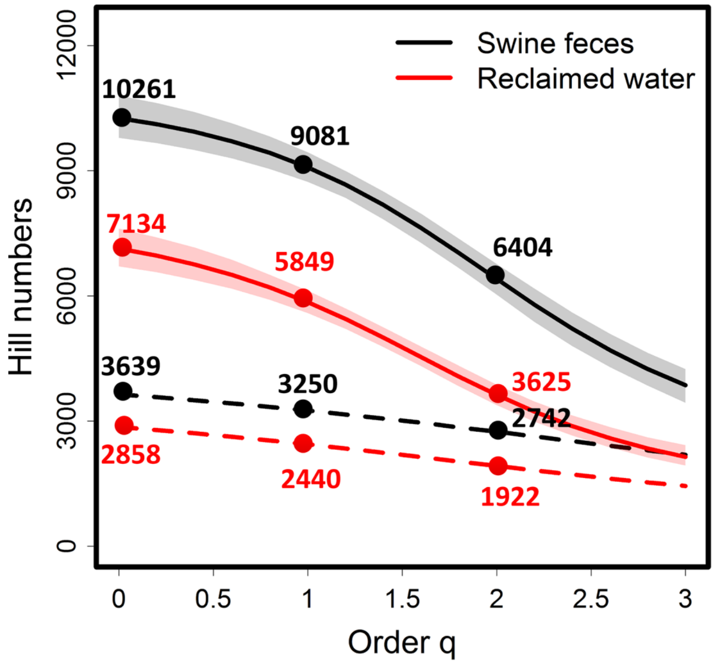

Figure 5 Asymptotic analysis: the asymptotic diversity profile as a function of order *q* based on the adjusted data. The empirical (dashed lines) and estimated (solid lines) diversity profiles for *q* between 0 and 3 based on the sample frequency counts of contig spectra from seven swine fecal viromes and the sample from four reclaimed fresh water viromes (*Allen et al., 2013*). Data are given in Table 1. The original singleton count is replaced by the estimated count given in Table 1. The plots for the swine feces sample are in black; the plots for the reclaimed water sample are in red. The 95% confidence intervals (shaded areas) are obtained by a bootstrap method based on 200 replications. The numbers (black for swine feces sample, and red for reclaimed water sample) show the empirical and estimated diversities for $q = 0$, 1 and 2.

CONCLUSION AND DISCUSSION

Whenever the singletons are uncertain or in doubt in sequencing data, it is worth applying our proposed estimator to estimate the singleton count; see Eq. (5). The discrepancy between our estimated singleton count and the observed count can be used to infer whether sequencing errors were present in data processing. Using the estimated number of singleton count and the original non-singleton frequency counts, we can quantify and compare microbial diversity for data sets with different sequencing error rates through non-asymptotic analysis (based on the plots of the sample-size- and coverage-based rarefaction and extrapolation sampling curves) and asymptotic analysis (based on the plot of a continuous asymptotic diversity profile estimator). Illustrative plots for sequencing data from viral metagenomes are shown in Fig. 4 (the non-asymptotic analysis) and Fig. 5 (the asymptotic analysis). Although we have focused on microbial data with spurious singleton counts, both our asymptotic and non-asymptotic approaches are also recommended for analyzing data with reliable singleton counts.

In highly diverse microbial communities, unless strong assumptions or parametric models are made, sampling data often do not provide sufficient information to accurately infer the number of undetected taxa in the sample. Thus, it is statistically infeasible to provide reliable estimates of taxa richness for the entire community. Our estimated species richness ($q = 0$ measure in our asymptotic analysis) theoretically is a lower bound. This implies that fair comparison of asymptotic species richness among multiple communities is not statistically feasible. In this case, fair comparison of taxa richness across multiple assemblages can be made by standardizing sample completeness (i.e., comparing taxa richness for a standardized fraction of population) based on coverage-based rarefaction and extrapolation sampling curves, as illustrated in the real data analysis. By contrast, when the diversity order q is away from 0 (say, $q \geq 1$), rare species have less impact on these diversities, and we generally can infer these diversities up to asymptotes and compare them across communities; see our illustrative example for interpretations. We recommend the use of an estimated diversity profile such as Fig. 5 for asymptotic analysis. If only one or two measures are desired in the inferences of highly diverse microbial diversity, then a perspective from Shannon diversity and Simpson diversity, instead of taxa richness, is more promising and more practical because we can accurately estimate Shannon and Simpson diversity not only for standardized samples but also their asymptotes. Besides, as shown in our simulation results (Figure S1 in Supplemental Text S2), the taxa richness estimator is seriously inflated or affected by spurious singleton counts, whereas the effect on Shannon diversity and Simpson diversity is less serious.

Our proposed estimator of singleton count is in terms of f_2, f_3 and f_4, provided these counts are reliable. A slight generalization of our method can be applied to estimate any frequency count. For example, supposing that singletons and doubletons are both uncertain, we can similarly derive an estimator of doubleton count based on f_3, f_4 and f_5 following exactly the same approach proposed in this paper. Subsequently, Eq. (5) then gives an estimate of singleton count based on the estimated doubleton count, f_3 and f_4. Consequently, our proposed non-asymptotic and asymptotic analyses can be similarly applied to data with the first two frequency counts being replaced by the estimated values. However, the sampling variance of the estimated diversity would be unavoidably increased.

In our approach, the original singleton count is discarded and replaced by our estimated count. In 16S rRNA sequencing or metagenomic sequencing, it is often standard practice to compare sequencing reads against a reference database, such as Greengenes, e.g., see *Turnbaugh et al. (2009)* or the software MOTHUR (*Schloss et al., 2009*). The Greengenes alignment tool helps adjust the original singleton count and alleviate the problem of sequencing error. Also, when there are multiple samples, singletons in a given sample have different probabilities of being spurious depending on their total number of reads across samples. Because many local singletons are not global singletons, the cross-sample information may also help adjust the original singleton count. Further investigation examining how to extend our framework to incorporate related covariates (such as cross-sample and database information) is merited.

Finally, we briefly discuss the phylogenetic diversity (PD) because of its broad interest and applications (*Martin, 2002*; *Lozupone & Knight, 2005*) in microbial studies. In this paper, all taxa are treated as if they were equally distinct and thus differences among sequences are not considered. *Faith's (1992)* PD is the most widely used PD metric to take into account phylogenetic differences among taxa. *Faith's (1992)* PD is defined as the total sum of branch lengths of a phylogenetic tree connecting all focal species. Based on sampling data, *Chao et al. (2015)* recently proposed a non-parametric estimator of the true PD (PD of the entire community, i.e., the observed PD in the sample plus the undetected PD). In the presence of sequencing errors, the inflated singleton count will also affect the estimation of the true PD. Since error-induced singletons will mostly likely fall in a closely related taxon, the effect may not be as pronounced as that in species richness estimation. More investigation is needed to tackle sequencing error and to adjust the *Chao et al. (2015)* PD estimator. Since *Faith's (1992)* PD does not incorporate taxa abundances, *Chao, Chiu & Jost (2010)* developed a class of abundance-sensitive PD measures which generalize *Faith's (1992)* PD to incorporate taxa abundances, and also extend Hill numbers to take into account phylogenetic relationships among taxa. How to extend the proposed analyses presented in this paper (the asymptotic and non-asymptotic analyses) to the class of abundance-sensitive PD is a worthwhile topic of future research.

ACKNOWLEDGEMENTS

The authors thank the Academic Editor (Rui Feng), Lou Jost and two anonymous reviewers for providing very helpful and thoughtful suggestions and comments, which have greatly improved the paper.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work is supported by the Taiwan Ministry of Science and Technology under Contract 104-2118-M-007-006-MY3 (for CHC) and 103-2628-M007-007 (for AC). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
Taiwan Ministry of Science and Technology: 104-2118-M-007-006-MY3 and 103-2628-M007-007.

Competing Interests

The authors declare that they have no competing interests.

Author Contributions

- Chun-Huo Chiu conceived and designed the experiments, performed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, prepared figures and/or tables, reviewed drafts of the paper, simulation study, theoretical derivation.
- Anne Chao conceived and designed the experiments, analyzed the data, contributed reagents/materials/analysis tools, wrote the paper, reviewed drafts of the paper, theoretical derivation.

Data Deposition

The following information was supplied regarding data availability:

The data are shown in Table 1 of the main text and in the Supplemental Information.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.1634#supplemental-information.

REFERENCES

Allen HK, Bunge J, Foster JA, Bayles DO, Stanton TB. 2013. Estimation of viral richness from shotgun metagenomes using a frequency count approach. *Microbiome* **1**:5 DOI 10.1186/2049-2618-1-5.

Allen HK, Looft T, Bayles DO, Humphrey S, Levine UY, Alt D, Stanton TB. 2011. Antibiotics in feed induce prophages in swine fecal microbiomes. *mBio* **2(6):**e00260–00211 DOI 10.1128/mBio.00260-11.

Bohannan BJ, Hughes J. 2003. New approaches to analyzing microbial biodiversity data. *Current Opinion in Microbiology* **6(3):**282–287 DOI 10.1016/S1369-5274(03)00055-9.

Buee M, Reich M, Murat C, Morin E, Nilsson RH, Uroz S, Martin F. 2009. 454 Pyrosequencing analyses of forest soils reveal an unexpectedly high fungal diversity. *New Phytologist* **184(2):**449–456 DOI 10.1111/j.1469-8137.2009.03003.x.

Bunge J, Böhning D, Allen H, Foster JA. 2012a. Estimating population diversity with unreliable low frequency counts. In: *Biocomputing 2012: Proceedings of the Pacific Symposium, Hackensack NJ.* Singapore: World Scientific Publication, 203–212.

Bunge J, Willis A, Walsh F. 2014. Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* **1**:427–445 DOI 10.1146/annurev-statistics-022513-115654.

Bunge J, Woodard L, Böhning D, Foster JA, Connolly S, Allen HK. 2012b. Estimating population diversity with CatchAll. *Bioinformatics* **28(7):**1045–1047 DOI 10.1093/bioinformatics/bts075.

Chao A. 1984. Nonparametric estimation of the number of classes in a population. *Scandinavian Journal of Statistics* **11(4):**265–270 DOI 10.2307/4615964.

Chao A. 1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* **43(4):**783–791 DOI 10.2307/2531532.

Chao A, Chiu CH, Hsieh T, Davis T, Nipperess DA, Faith DP. 2015. Rarefaction and extrapolation of phylogenetic diversity. *Methods in Ecology and Evolution* **6(4):**380–388 DOI 10.1111/2041-210X.12247.

Chao A, Chiu C-H, Jost L. 2010. Phylogenetic diversity measures based on Hill numbers. *Philosophical Transactions of the Royal Society B: Biological Sciences* **365(1558):**3599–3609 DOI 10.1098/rstb.2010.0272.

Chao A, Chiu C-H, Jost L. 2014. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. *Annual Review of Ecology, Evolution, and Systematics* **45**:297–324 DOI 10.1146/annurev-ecolsys-120213-091540.

Chao A, Gotelli NJ, Hsieh T, Sander EL, Ma K, Colwell RK, Ellison AM. 2014. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* **84(1):**45–67 DOI 10.1890/13-0133.1.

Chao A, Jost L. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* **93(12):**2533–2547 DOI 10.1890/11-1952.1.

Chao A, Jost L. 2015. Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution* **6(8):**873–882 DOI 10.1111/2041-210X.12349.

Chao A, Wang Y, Jost L. 2013. Entropy and the species accumulation curve: a novel entropy estimator via discovery rates of new species. *Methods in Ecology and Evolution* **4(11):**1091–1100 DOI 10.1111/2041-210X.12108.

Chiu CH, Wang YT, Walther BA, Chao A. 2014. An improved nonparametric lower bound of species richness via a modified Good–Turing frequency formula. *Biometrics* **70(3):**671–682 DOI 10.1111/biom.12200.

Colwell RK, Chao A, Gotelli NJ, Lin S-Y, Mao CX, Chazdon RL, Longino JT. 2012. Models and estimators linking individual-based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* **5(1):**3–21 DOI 10.1093/jpe/rtr044.

Colwell RK, Coddington JA. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B: Biological Sciences* **345(1311):**101–118 DOI 10.1098/rstb.1994.0091.

Curtis TP, Sloan WT, Scannell JW. 2002. Estimating prokaryotic diversity and its limits. *Proceedings of the National Academy of Sciences of the United States of America* **99(16):**10494–10499 DOI 10.1073/pnas.142680199.

Dickie IA. 2010. Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytologist* **188(4):**916–918 DOI 10.1111/j.1469-8137.2010.03473.x.

Doll HM, Armitage DW, Daly RA, Emerson JB, Goltsman DS, Yelton AP, Kerekes J, Firestone MK, Potts MD. 2013. Utilizing novel diversity estimators to quantify multiple dimensions of microbial biodiversity across domains. *BMC Microbiology* **13:**259 DOI 10.1186/1471-2180-13-259.

Ellison AM. 2010. Partitioning diversity. *Ecology* **91(7):**1962–1963 DOI 10.1890/09-1692.1.

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* **61(1):**1–10 DOI 10.1016/0006-3207(92)91201-3.

Fierer N, Hamady M, Lauber CL, Knight R. 2008. The influence of sex, handedness, and washing on the diversity of hand surface bacteria. *Proceedings of the National Academy of Sciences* **105(46):**17994–17999 DOI 10.1073/pnas.0807920105.

Good IJ. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika* **40(3–4):**237–264 DOI 10.1093/biomet/40.3-4.237.

Good IJ. 1983. *Good Thinking: The Foundations of Probability and Its Applications.* Minneapolis: University of Minnesota Press.

Good IJ. 2000. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma. *Journal of Statistical Computation and Simulation* **66(2):**101–111 DOI 10.1080/00949650008812016.

Good IJ, Toulmin G. 1956. The number of new species and the increase of population coverage when a sample is increased. *Biometrika* **43(1–2):**45–63 DOI 10.1093/biomet/43.1-2.45.

Gotelli N, Chao A. 2013. Measuring and estimating species richness, species diversity, and biotic similarity from sampling data. In: Levin SA, eds. *Encyclopedia of Biodiversity.* Waltham: Academic, 195–211 DOI 10.1016/B978-0-12-384719-5.00424-X.

Haegeman B, Hamelin J, Moriarty J, Neal P, Dushoff J, Weitz JS. 2013. Robust estimation of microbial diversity in theory and in practice. *The ISME Journal* **7:**1092–1101 DOI 10.1038/ismej.2013.10.

Haegeman B, Sen B, Godon J-J, Hamelin J. 2014. Only simpson diversity can be estimated accurately from microbial community fingerprints. *Microbial Ecology* **68(2):**169–172 DOI 10.1007/s00248-014-0394-5.

Haegeman B, Vanpeteghem D, Godon JJ, Hamelin J. 2008. DNA reassociation kinetics and diversity indices: richness is not rich enough. *Oikos* **117(2):**177–181 DOI 10.1111/j.2007.0030-1299.16311.x.

Hill M. 1973. Diversity and evenness: a unifying notation and its consequences. *Ecology* **54(2):**427–432 DOI 10.2307/1934352.

Hill TC, Walsh KA, Harris JA, Moffett BF. 2003. Using ecological diversity measures with bacterial communities. *FEMS Microbiology Ecology* **43(1):**1–11 DOI 10.1111/j.1574-6941.2003.tb01040.x.

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. 2001. Counting the uncountable: statistical approaches to estimating microbial diversity. *Applied and Environmental Microbiology* **67(10):**4399–4406 DOI 10.1128/AEM.67.10.4399-4406.2001.

Huse SM, Welch DM, Morrison HG, Sogin ML. 2010. Ironing out the wrinkles in the rare biosphere through improved OTU clustering. *Environmental Microbiology* **12(7):**1889–1898 DOI 10.1111/j.1462-2920.2010.02193.x.

Jost L. 2006. Entropy and diversity. *Oikos* **113(2):**363–375 DOI 10.1111/j.2006.0030-1299.14714.x.

Jost L. 2007. Partitioning diversity into independent alpha and beta components. *Ecology* **88(10):**2427–2439 DOI 10.1890/06-1736.1.

Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental Microbiology* **12(1):**118–123 DOI 10.1111/j.1462-2920.2009.02051.x.

Lozupone K, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology* **71(12):**8228–8235 DOI 10.1128/AEM.71.12.8228-8235.2005.

MacArthur RH. 1965. Patterns of species diversity. *Biological Reviews* **40(4):**510–533 DOI 10.1111/j.1469-185X.1965.tb00815.x.

Martin AP. 2002. Phylogenetic approaches for describing and comparing the diversity of microbial communities. *Applied and Environmental Microbiology* **68(8):**3673–3682 DOI 10.1128/AEM.68.8.3673-3682.2002.

Øvreås L, Curtis T. 2011. Microbial diversity and ecology. In: Magurran AE, McGill BJ, eds. *Biological Diversity: Frontiers in Measurement and Assessment.* Oxford: Oxford University Press, 221–236.

Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT. 2009. Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods* **6:**639–641 DOI 10.1038/nmeth.1361.

Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. 2011. Removing noise from pyrosequenced amplicons. *BMC Bioinformatics* **12:**38 DOI 10.1186/1471-2105-12-38.

Robbins HE. 1968. Estimating the total probability of the unobserved outcomes of an experiment. *The Annals of Mathematical Statistics* **39(1):**256–257 DOI 10.1214/aoms/1177698526.

Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SU, Camargo FAO, Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *The ISME Journal* **1:**283–290 DOI 10.1038/ismej.2007.53.

Schloss PD, Handelsman J. 2005. Introducing DOTUR, A computer program for defining operational taxonomic units and estimating species richness. *Applied and Environmental Microbiology* **71(3):**1501–1506 DOI 10.1128/AEM.71.3.1501-1506.2005.

Schloss PD, Handelsman J. 2006. Toward a census of bacteria in soil. *PLoS Computational Biology* **2(7):**e92 DOI 10.1371/journal.pcbi.0020092.

Schloss PD, Handelsman J. 2008. A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics* **9**:34 DOI 10.1186/1471-2105-9-34.

Schloss PD, Westcott SL, Ryabin T, Hall JR, Hartmann M, Hollister EB, Lesniewski RA, Oakley BB, Parks DH, Robinson CJ. 2009. Introducing mothur: open-source, platform-independent, community-supported software for describing and comparing microbial communities. *Applied and Environmental Microbiology* **75(23):**7537–7541 DOI 10.1128/AEM.01541-09.

Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere". *Proceedings of the National Academy of Sciences of the United States of America* **103(32):**12115–12120 DOI 10.1073/pnas.0605127103.

Tóthmérész B. 1995. Comparison of different methods for diversity ordering. *Journal of Vegetation Science* **6(2):**283–290 DOI 10.1234/12345678.

Turnbaugh PJ, Gordon JI. 2009. The core gut microbiome, energy balance and obesity. *Journal of Physiology* **587(17):**4153–4158 DOI 10.1113/jphysiol.2009.174136.

Turnbaugh PJ, Hamady M, Yatsunenko T, Cantarel BL, Duncan A, Ley RE, Sogin ML, Jones WJ, Roe BA, Affourtit JP, Egholm M, Henrissat B, Heath AC, Knight R, Gordon JI. 2009. A core gut microbiome in obese and lean twins. *Nature* **457(7228):**480–484 DOI 10.1038/nature07540.

Willis A, Bunge J. 2015. Estimating diversity via frequency ratios. *Biometrics* **71(4):**1042–1049 DOI 10.1111/biom.12332.