A peer-reviewed version of this preprint was published in PeerJ on 1 February 2016.

Chiu C, Chao A. 2016. Estimating and comparing microbial diversity in the presence of sequencing errors. PeerJ 4:e1634
https://doi.org/10.7717/peerj.1634

Estimating and comparing microbial diversity in the presence of sequencing errors

Chun-Huo Chiu, Anne Chao

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 for the first time a nonparametric estimator of the true singleton count to replace the spurious singleton count. Our estimator of the true singleton count is in terms of the frequency counts of doubletons, tripletons and quadrupletons. To quantify microbial diversity, 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), and Simpson diversity (q=2). Based on the estimated singleton count and the original non-singleton frequency counts, two statistical approaches are developed to compare microbial diversity for multiple communities. (1) A non-asymptotic approach based on standardizing sample size or sample completeness via seamless rarefaction and extrapolation sampling curves of Hill numbers. (2) An asymptotic approach based on a continuous diversity (Hill number) profile which depicts the estimated asymptotes of diversities as a function of order q. Replacing the spurious singleton count by our estimated count, we can greatly remove the positive biases associated with diversity estimates due to spurious singletons in the two approaches and make fair comparison across microbial communities, as illustrated in applying our method to analyze sequencing data from viral metagenomes.

Estimating and comparing microbial diversity in the presence of sequencing errors

Chun-Huo Chiu[1] and Anne Chao[1,*]

[1] Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan, 30043

*Corresponding author: Anne Chao, Institute of Statistics, National Tsing Hua University, Hsin-Chu, Taiwan, 30043

E-mail: chao@stat.nthu.edu.tw

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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 for the first time a nonparametric estimator of the true singleton count to replace the spurious singleton count. Our estimator of the true singleton count is in terms of the frequency counts of doubletons, tripletons and quadrupletons. To quantify microbial diversity, 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), and Simpson diversity (q=2). Based on the estimated singleton count and the original non-singleton frequency counts, two statistical approaches are developed to compare microbial diversity for multiple communities. (1) A non-asymptotic approach based on standardizing sample size or sample completeness via seamless rarefaction and extrapolation sampling curves of Hill numbers. (2) An asymptotic approach based on a continuous diversity (Hill number) profile which depicts the estimated asymptotes of diversities as a function of order q. Replacing the spurious singleton count by our estimated count, we can greatly remove the positive biases associated with diversity estimates due to spurious singletons in the two approaches and

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

make fair comparison across microbial communities, as illustrated in applying our method to analyze sequencing data from viral metagenomes.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

INTRODUCTION

Advances in high-throughput DNA sequencing have opened a novel way to assess hyper-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, 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. 2012; Bunge et al. 2012). 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 the older 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; 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, 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 examples, Haegeman et al. (2008; 2013; 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. However, in practice, community parameters are unknown and thus the true diversity (i.e., asymptotic diversity) must be estimated from sampling data and statistical methods are required. See the asymptotic analysis in later text.

In this paper, we propose two statistical approaches to make fair comparisons of microbial diversity across multiple communities. Our first approach is a non-asymptotic approach based on standardizing sample size or sample completeness (as measured by sample coverage; see below) via an integrated rarefaction and extrapolation curve. In this approach, the diversities of multiple communities can be compared for standardized finite sample sizes or standardized sample overages. 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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, $0 \le q < 2$) heavily depend on the low frequency counts especially singletons (Chao et al., 2014).

Our second approach is an asymptotic approach based on a continuous diversity profile which depicts the estimated asymptotes of diversities as a function of order q. 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 and Jost (2015) recently proposed a resolution to obtain a diversity profile estimator, which infers the asymptotes of diversities, i.e., the diversity when the sample size tends to infinity or sample completeness tends to unity. The negative bias associated with the empirical diversity curve due to undetected species can be greatly reduced. They 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 and Jost's (2015) diversity profile estimator for low value of q ($0 \le q < 2$) is strongly affected by the low frequency counts. This is mainly because the observed rare species that

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

produce low frequencies carry nearly all 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 thus are 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 described above, our suggested approaches will be seriously affected if the inflated singleton count is not corrected. 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 were also recently proposed to correct the number of spurious singletons and estimate diversity. Bunge et al. (2012; 2014) proposed a parametric mixture model and a method using "left-censored" data; Willis and Bunge (2014) proposed an approach using the ratio of two successive frequency counts. These pioneering 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 due to limited microbial data, and different communities may not be compared due to different parametric assumptions.

In this paper, we propose for the first time a novel nonparametric approach to estimate the true number of singletons in the presence of sequencing errors. We derive here 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

developed by the founder of modern computer science Alan Turing, and I. J. Good (1953; 2000). Our estimator of singleton count is thus 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 sequencing errors were present or not 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 possibly spurious singletons. 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. Sequencing data from viral metagenomes (Allen et al., 2011; Allen et al., 2013) are used for illustration. 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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, $i = 1, 2, ..., S$, $\sum_{i=1}^{S} p_i = 1$, and X_i be the number of individual of ith species detected in the sample of size n. Let f_k (abundance frequency counts), $k = 1, 2, ..., n$, 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 $\{p_1, p_2, ..., p_S\}$, the Hill number of order q is defined as:

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

The measure for q=0 counts species equally without regard to their relative abundances. 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 and 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 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 $0 \leq q \leq 3$.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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; see (Good, 1953; Good & Toulmin, 1956; Good, 2000). The Good–Turing frequency theory can be formulated as follows: For those species that appeared r times, $r = 0, 1, \ldots$, in a sample of size n, how one can 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

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

Chiu et al. (2014) modified the Good–Turing estimator to obtain a more accurate formula:

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

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

Singleton count estimation

In the Chao1 lower bound of species richness (1984), the zero-frequency count is estimated by the frequencies of singletons and doubletons. 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

formula for the k-th frequency count is:

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

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 Equation (4b) as

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

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

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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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

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

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

The right hand side of the above formula will be positive for reasonably large sample size, 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 (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:

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

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:

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

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

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

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).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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 as described below to compare non-asymptotic diversities across multiple assemblages. 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 in Equation (7).

(2) Coverage-based rarefaction and extrapolation up to a maximum coverage. Chao and 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; 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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 in Equation (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 R package "iNEXT" which can be also downloaded from Anne Chao's website http://chao.stat.nthu.edu.tw/blog/software-download/.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

Asymptotic approach: diversity profile estimation based on adjusted data

The Chao and Jost (2015) diversity profile estimator based on the adjusted singleton count $\hat{f}_1$ and the original non-singleton frequency counts can be expressed as

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

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

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

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 estimator of order q in each profile represents the asymptote in the rarefaction and extrapolation curves described above. To obtain the profile estimator 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 Equation (7) reduces to the adjusted Chao1 estimator given in Equation (6a). Thus, it is generally a minimum number of species and thus cannot be used for ranking or comparing multiple communities. For q=1, the estimation of the Shannon diversity

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

297 from incomplete samples is surprisingly nontrivial and has been extensively discussed in many
298 research fields; see (Chao, Wang & Jost, 2013) for a review and a low-bias estimator. The
299 estimator (7) for q=1 reduces to their Shannon diversity estimator (given below), which can be
300 compared across communities.

301 $$ {}^{1}\hat{D}_{adj} = \exp\left(\sum_{1\le X_i\le n-1} \frac{X_i}{n}\left(\sum_{k=X_i}^{n-1} \frac{1}{k} \right) + \frac{\hat{f}_1}{n}(1-A)^{-n+1}\left[-\log A - \sum_{r=1}^{n-1} \frac{(1-A)^r}{r} \right] \right) . $$

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

307 $$ {}^{2}\hat{D}_{adj} = \left(\sum_{X_i\ge 2} \frac{X_i(X_i-1)}{n(n-1)} \right)^{-1} . $$

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

311

312 # SIMULATION RESULTS

313 Since both non-asymptotic and asymptotic analyses depend on the quality of the estimated
314 singleton count, it is essential to investigate the performance of the proposed estimator in Equation
315 (5). We conducted a simulation by generating data from six species abundance distributions with
316 various degrees of heterogeneity in species relative abundances (details are provided in
317 Supplemental Text S2). In each model, we fixed the number of species at S=2000 to mimic
318 microbial communities. Then for each given model, we considered a range of sample sizes (n =

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

2000 to 10000 in an increment of 2000).

For each combination of abundance model and sample size, we generated two types of data: (i) true data without sequencing errors, and (ii) spurious data with a sequencing error rate of 10%, i.e., there was 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 1000 simulation trials) of three singleton counts as a function of sample size. The three singleton counts include those obtained from the true data, spurious data, and our proposed estimation method. The pattern revealed by these plots is summarized below.

Fig. 1 reveals that the number of singletons for the true data (dotted curve in each panel) generally declines with sample size when sample size becomes sufficiently large, whereas the number of singletons for spurious data (dashed 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 our estimator of singleton count generally matches closely the true number of singletons (solid line in each panel), 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 dashed curves represent the singleton counts for data), our estimator can greatly reduce the raw singleton count and make proper correction. 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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 estimated asymptotes of diversities are provided in Supplemental Text S2.

APPLICATION RESULTS

We next present the 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 and estimated diversities are shown in Table 2.

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

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 Equation (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).

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	9988	4974	8025	2831	605	129	41	16	8	4	2	1	1	1	0	0
Reclaimed water	9973	4092	7986	2105	518	129	50	24	12	7	5	3	2	1	1	1

Table 2. Empirical diversities and the estimated asymptotes of diversities for the phage metagenomic data given in Table 1. CI = confidence interval. The estimated asymptotes are computed from the adjusted data (i.e., the original data with the observed singleton count being replaced by the estimated value given in Table 1)

Sample	Diversity	Original empirical diversity	Adjusted empirical diversity	Estimated asymptote of diversity	SE	95% lower CI	95% upper CI
Swine feces	Species richness ($q = 0$)	8833	3639	10261	376	9565	11039
	Shannon diversity ($q = 1$)	8289	3250	9081	203	8684	9479
	Simpson diversity ($q = 2$)	7348	2742	6404	180	6051	6757
Reclaimed water	Species richness ($q = 0$)	8739	2858	7134	273	6632	7703
	Shannon diversity ($q = 1$)	8066	2440	5849	130	5595	6104
	Simpson diversity ($q = 2$)	6817	1922	3625	116	3398	3852

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

In the swine feces original data, there were 8833 taxa among 9988 individuals (sequences); the number of singletons was f_1 =8025, and the number of doubletons was f_2 =605. In the reclaimed water data, there were 8739 taxa among 9973 individuals, and the first two frequency counts are f_1 =7986 and f_2 =518. In these two original samples, most of the frequencies are concentrated on singletons. Using Equation (5), we obtain an estimated singleton count 2831 for swine feces sample, and 2105 for reclaimed water sample. Thus, the adjusted sample sizes are declined to 4974 and 4092 respectively. For each sample, the estimated singleton count is substantially less than the observed singleton count, revealing sequencing errors were present. Consequently, the Chao1 lower bounds 62057 and 70299 respectively for the original data are greatly inflated due to spurious singletons. All the following analyses are based on the adjusted data, unless otherwise stated.

In Fig. 2, we plot 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 10000 (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. 2 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. 3 the sample-size- and coverage-based rarefaction and extrapolation curves along with 95% confidence intervals in Fig. 3 for three measures: q=0, 1 and 2. The sample-size-based sampling curve is extrapolated up to a maximum size of 10000, whereas the coverage-based sampling curve is extended up to the coverage of the

size 10000, 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. 3 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, signifying significant difference. As stated earlier, the extrapolation for Shannon and Simpson diversity, but rarely species richness, can often be reliably extended to infinity or complete coverage to reach the asymptotic diversity estimate. Therefore, for Shannon diversity (common taxa richness) and Simpson diversity (dominant taxa richness), data conclude 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. 3, but also for entire viromes. (This is also supported by the asymptotic analysis below.) For taxa richness, data support the conclusion up to a standardized 62.9% fraction of each virome (the upper right panel in Fig. 3). Beyond that, 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. 4 when q is between 0 and 3. The estimated asymptotes of diversities for the special cases of q= 0, 1 and 2 are shown in Table 2 and also shown next to an arrow at the right-hand end of each rarefaction/extrapolation plot in Fig. 3. The empirical diversities based on the original spurious data and for the adjusted data (Table 2 and Fig. 4) imply that the two viromes have limited difference in each of the three measures. In contrast, the plots in Fig. 4 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

the Simpson diversity, confirming our earlier statement in the preceding paragraph. Table 2 and Fig. 4 show that the adjusted Chao1 estimator in Equation (6a) gives an estimate of 10261 taxa for swine feces and 7134 taxa for reclaimed water virome. Each is five times that obtained from CatchAll (Allen et al., 2013). As discussed earlier, since the adjusted Chao1 estimate represents only minimum richness, it cannot be used to rank the taxa richness of the two entire viromes. Similarly, for any value q close to zero, our estimated asymptotes also represent lower bounds only. So we generally cannot compare the estimated low-order asymptotes of diversities including taxa richness across multiple communities; see the next section for more discussions. In Supplemental Table S1, we also give all the estimated asymptotes of diversities for other data sets provided in Allen et al. (2013).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

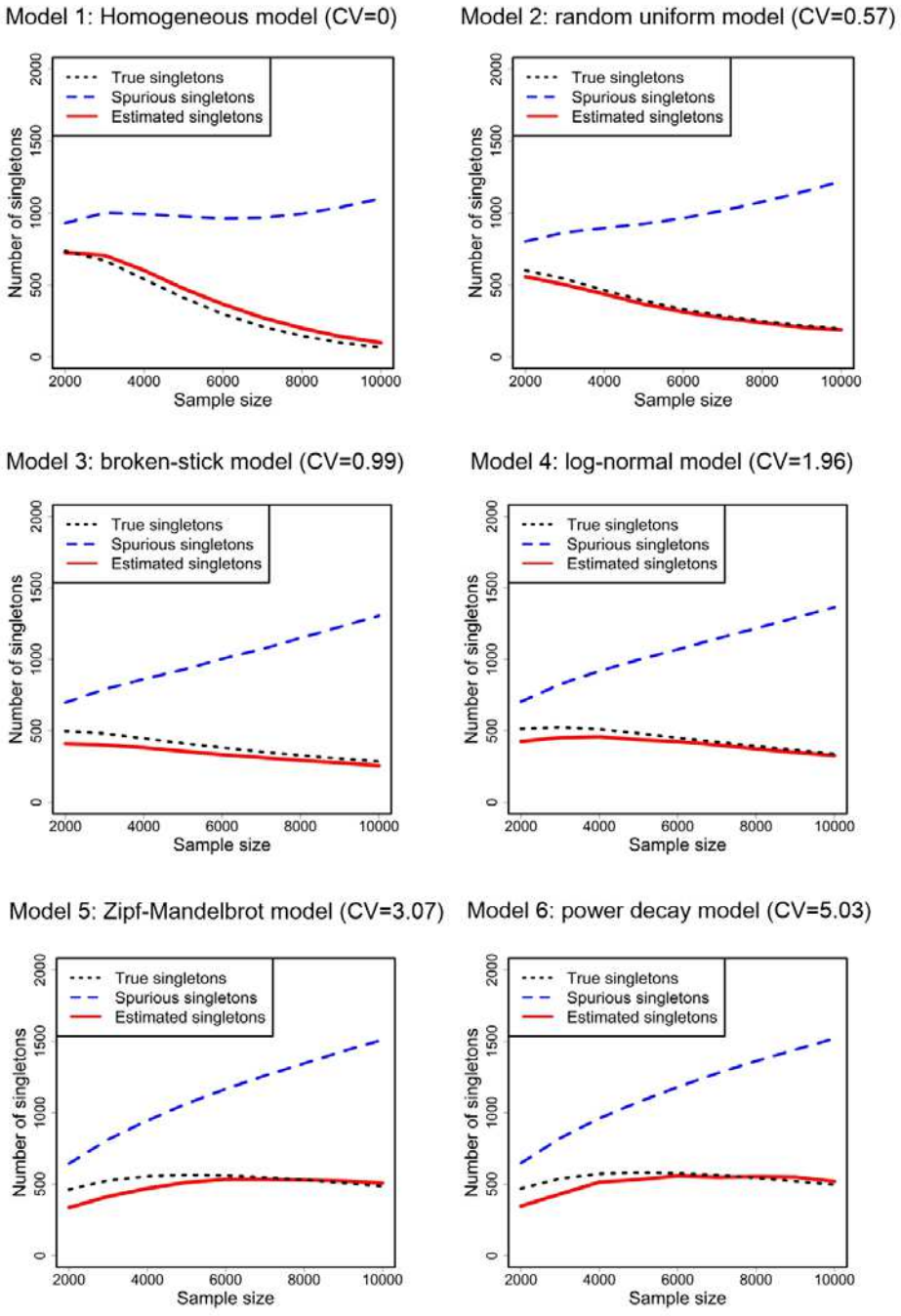

Fig 1. Plots of the average values of three singleton counts as a function of sample size. The three singleton counts include those obtained from the true data, spurious data, and the estimated method based on Equation (5). All values are averaged over 1000 simulation trials under six species abundance 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.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

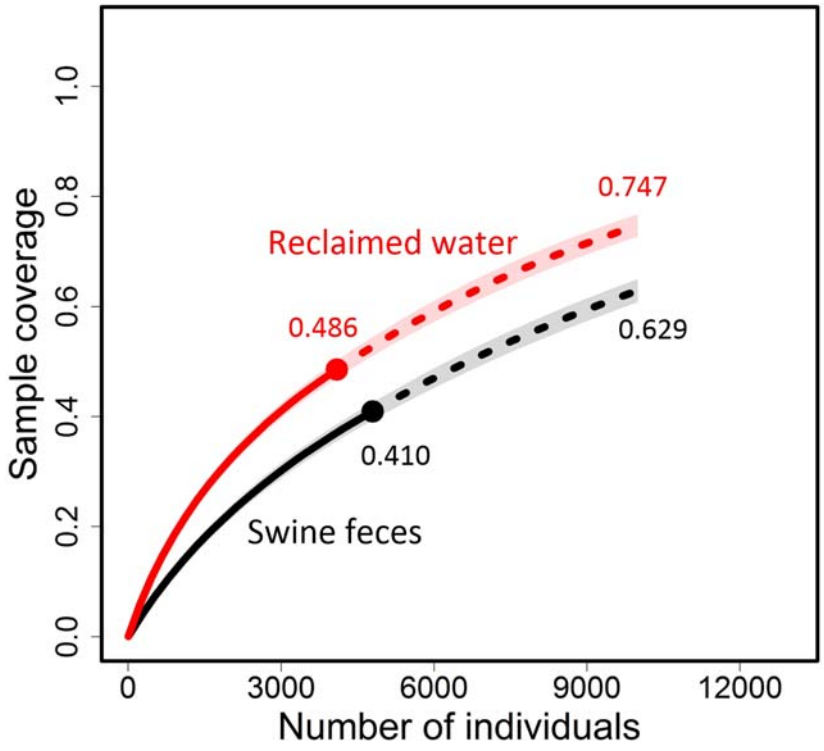

Fig 2. The sample completeness curve based on the adjusted data.

Plots of sample coverage for rarefied samples (solid line) and extrapolated samples (dashed line) as a function of 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 10000, 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 10000.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

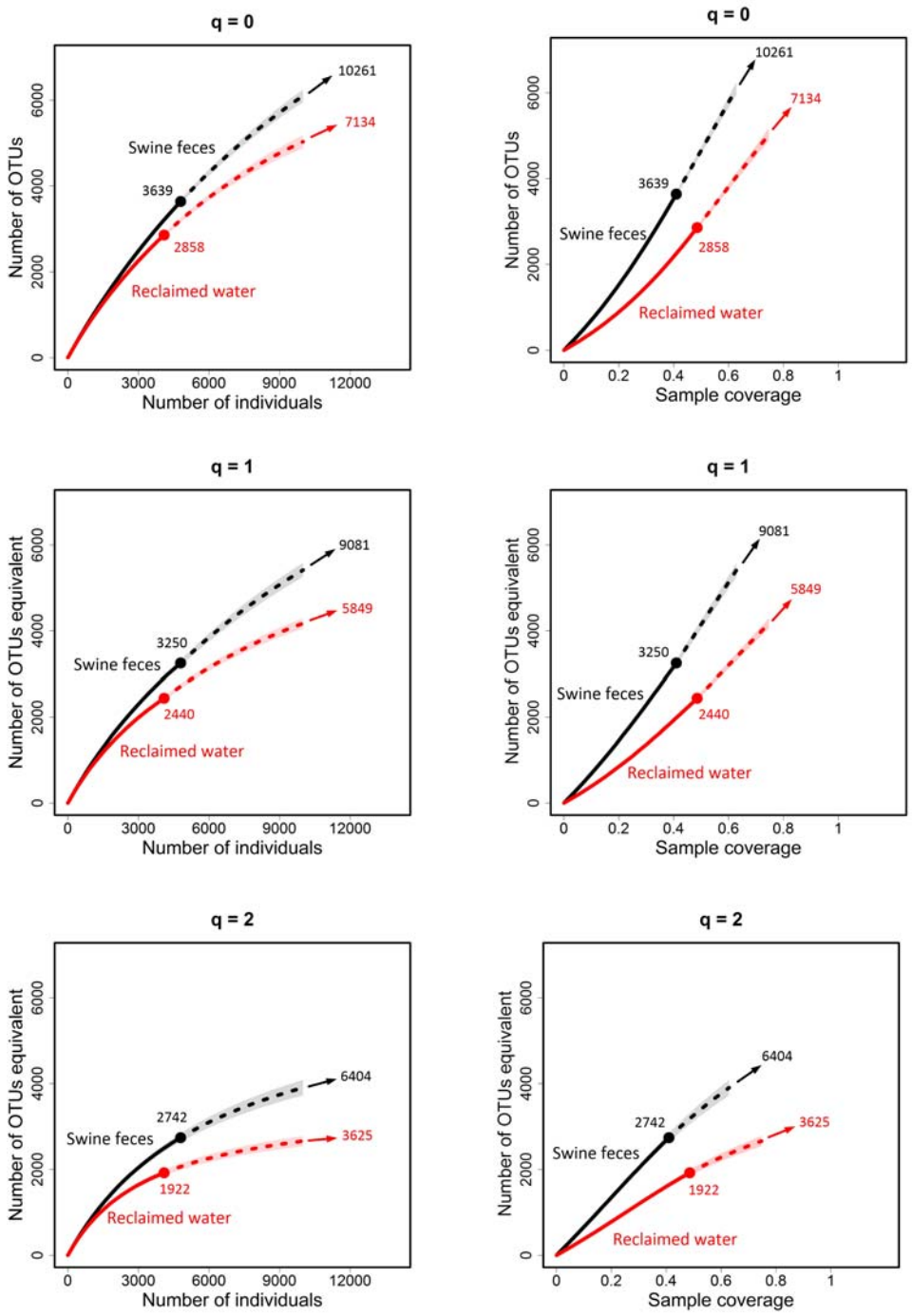

Fig 3. Non-asymptotic analysis: the rarefaction and extrapolation sampling curves based on the adjusted data. Comparison of sample-size-based (left panels) and sample-coverage-based (right panels) rarefaction and extrapolation for species richness (upper panels), Shannon diversity (middle panels) and Simpson diversity (lower panels) 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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 10000. 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. 2). 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.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

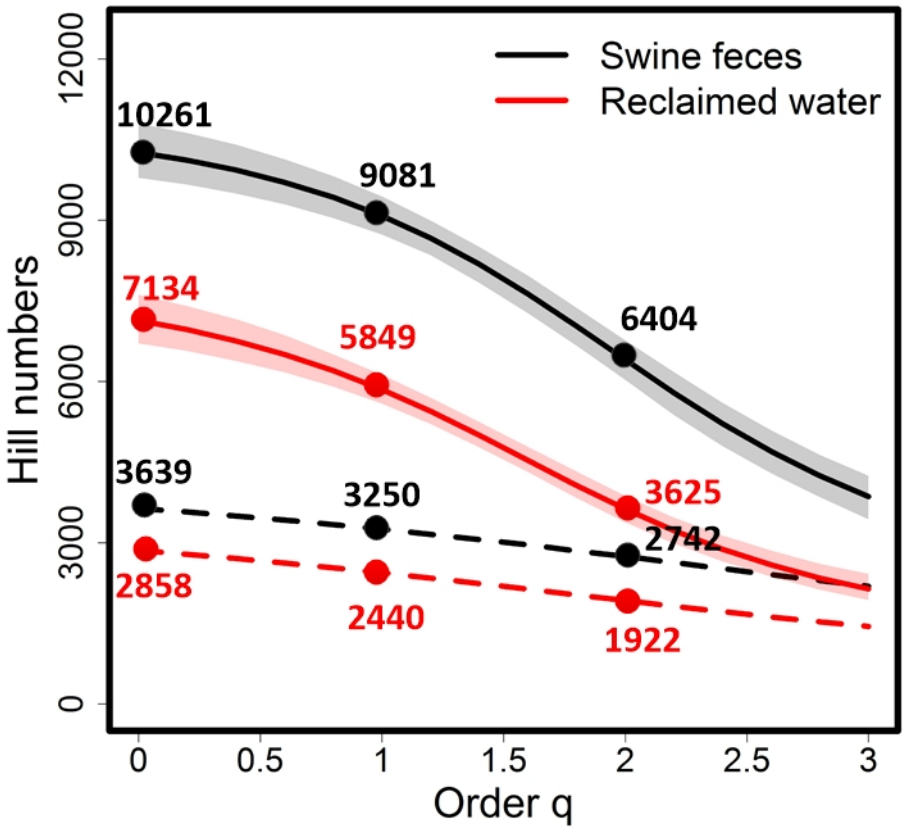

Fig 4: Asymptotic analysis: the asymptotic diversity profile as a function of order *q* based on the adjusted data. The empirical (dashed lines) and estimated (solid lines) diversity profiles for *q* between 0 and 3 based on the sample frequency counts of contig spectra from seven swine fecal viromes 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.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

CONCLUSION AND DISCUSSION

Whenever the singletons are uncertain or in doubt in sequencing data, it is worth applying our proposed estimator of the true singleton count; see Equation (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 by 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. 3 (the non-asymptotic analysis) and Fig. 4 (the asymptotic analysis).

In hyper-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 in 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 species richness among multiple communities is not statistically feasible. In this case, fair comparison of taxa richness across multiple assemblages can only 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. However, 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. Thus, in the inferences of hyper-diverse microbial diversity, a perspective from

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

Shannon diversity and Simpson diversity, instead of taxa richness, is more promising and more practical because accurate estimation of taxa richness is almost unattainable.

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, suppose 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, Equation (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.

Finally, we briefly discuss the phylogenetic diversity (PD) because of its broad interest and applications (Mattin, 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 PD (1992) is the most widely used PD metric to take into account phylogenetic differences among taxa. Faith's PD is defined as 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 un-detected PD). When sequencing error is present, the inflated singleton count will also seriously affect the estimation. More investigation is needed to tackle sequencing error and to adjust the PD estimator. Since Faith's PD does not incorporate taxa abundances, Chao, Chiu and Jost (2010) developed a class of abundance-sensitive PD measures which generalize Faith's 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

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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 Lou Jost for editing an earlier version and providing helpful and thoughtful suggestions and comments. CHC is supported by a post-doctoral fellowship, National Tsing Hua University, Taiwan. This research is supported by Taiwan Ministry of Science and Technology under Contract 103-2628-M007-007.

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:e00260–00211. DOI: 10.1128/mBio.00260-11.

Bohannan BJ, Hughes J. 2003. New approaches to analyzing microbial biodiversity data. *Current Opinion in Microbiology* 6:282–287. 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* 184:449–456. DOI: 10.1111/j.1469-8137.2009.03003.x.

Bunge J, Böhning D, Allen H, Foster JA. 2012. Estimating population diversity with unreliable low frequency counts. In Biocomputing 2012: Proceedings of the Pacific Symposium, Hackensack, NJ: 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. 2012. Estimating population diversity with CatchAll. *Bioinformatics* 28:1045–1047. DOI: 10.1093/bioinformatics/bts075.

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

Chao A.1987. Estimating the population size for capture-recapture data with unequal catchability. *Biometrics* 43:783–791. 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* 6:380–388. 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* 365:3599–3609. 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* 84:45–67. DOI:10.1890/13-0133.1.

Chao A, Jost L. 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* 93:2533–2547. 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* 6:873–882. DOI: 10.1111/2041-210X.12349.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

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* 4:1091–1100. 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* 70:671–682. 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* 5:3–21. DOI: 10.1093/jpe/rtr044.

Colwell RK, Coddington JA. 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society B: Biological Sciences* 345:101–118. 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 USA* 99:10494–10499. DOI: 10.1073/pnas.142680199.

Dickie IA. 2010. Insidious effects of sequencing errors on perceived diversity in molecular surveys. *New Phytologist* 188:916–918. 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 91:1962–1963. DOI: 10.1890/09-1692.1.

Faith DP. 1992. Conservation evaluation and phylogenetic diversity. *Biological Conservation* 61:1–10. 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*

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

105:17994–17999. DOI: 10.1073/pnas.0807920105.

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

Good IJ. 2000. Turing's anticipation of empirical Bayes in connection with the cryptanalysis of the naval Enigma. *Journal of Statistical Computation and Simulation* 66:101–111. 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* 43:45–63. 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, ed. *Encyclopedia of Biodiversity*. Waltham: Academic, 195–211.

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* 68:169–172. 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* 117:177–181. DOI: 10.1111/j.2007.0030-1299.16311.x.

Hill M. 973. Diversity and evenness: A unifying notation and its consequences. *Ecology* 54:427–432. DOI: 10.2307/1934352.

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

Hughes JB, Hellmann JJ, Ricketts TH, Bohannan BJ. 2001. Counting the uncountable: statistical

671 approaches to estimating microbial diversity. *Applied and Environmental Microbiology*
672 67:4399–4406. DOI: 10.1128/AEM.67.10.4399-4406.2001.
673
674 Huse SM, Welch DM, Morrison HG, Sogin ML. 2010. Ironing out the wrinkles in the rare
675 biosphere through improved OTU clustering. *Environmental Microbiology* 12:1889–1898.
676 DOI: 10.1111/j.1462-2920.2010.02193.x.
677
678 Jost L. 2006. Entropy and diversity. *Oikos* 113:363–375. DOI: 10.1111/j.2006.0030-1299.14714.x.
679
680 Jost L. 2007. Partitioning diversity into independent alpha and beta components. *Ecology*
681 88:2427–2439. DOI: 10.1890/06-1736.1.
682
683 Kunin V, Engelbrektson A, Ochman H, Hugenholtz P. 2010. Wrinkles in the rare biosphere:
684 pyrosequencing errors can lead to artificial inflation of diversity estimates. *Environmental*
685 *Microbiology* 12:118–123. DOI: 10.1111/j.1462-2920.2009.02051.x.
686
687 Logares R, Haverkamp THA, Kumar S, Lanzén A, Nederbragt AJ, Quince C, Kauserud H. 2012.
688 Environmental microbiology through the lens of high-throughput DNA sequencing: synopsis of
689 current platforms and bioinformatics approaches. *Journal of Microbiological Methods* 91:106–113.
690 DOI:10.1016/j.mimet.2012.07.017.
691
692 Lozupone K, Knight R. 2005. UniFrac: a new phylogenetic method for comparing microbial
693 communities. *Applied and Environmental Microbiology* 71:8228–8235. DOI:
694 10.1128/AEM.71.12.8228-8235.2005.
695
696 MacArthur RH. 1965. Patterns of species diversity. *Biological Reviews* 40:510–533. DOI:
697 10.1111/j.1469-185X.1965.tb00815.x
698
699 Martin AP. 2002. Phylogenetic approaches for describing and comparing the diversity of microbial
700 communities. *Applied and Environmental Microbiology* 68:3673–3682. DOI:
701 10.1128/AEM.68.8.3673-3682.2002.
702
703 Øvreås L, Curtis T. 2011. Microbial diversity and ecology. In: Magurran, AE and McGill, BJ, eds.
704 *Biological diversity: frontiers in measurement and assessment*. Oxford: Oxford University Press:
705 Oxford, 221–236.
706
707 Quince C, Lanzén A, Curtis TP, Davenport RJ, Hall N, Head IM, Read LF, Sloan WT. 2009.

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015

708 Accurate determination of microbial diversity from 454 pyrosequencing data. *Nature Methods*
709 6:639–641. DOI:10.1038/nmeth.1361.
710
711 Quince C, Lanzen A, Davenport RJ, Turnbaugh PJ. 2011. Removing noise from pyrosequenced
712 amplicons. *BMC Bioinformatics* 12:38. DOI:10.1186/1471-2105-12-38.
713
714 Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SU, Camargo FAO,
715 Farmerie WG, Triplett EW. 2007. Pyrosequencing enumerates and contrasts soil microbial
716 diversity. *The ISME Journal* 1:283–290. DOI:10.1038/ismej.2007.53.
717
718 Schloss PD, Handelsman J. 2005. Introducing DOTUR, A computer program for defining
719 operational taxonomic units and estimating species richness. *Applied and Environmental*
720 *Microbiology* 71:1501–1506. DOI: 10.1128/AEM.71.3.1501-1506.2005.
721
722 Schloss PD, Handelsman J. 2006. Toward a census of bacteria in soil. *PLoS Computational*
723 *Biology* 2: e92. DOI: 10.1371/journal.pcbi.0020092.
724
725 Schloss PD, Handelsman J. 2008. A statistical toolbox for metagenomics: assessing functional
726 diversity in microbial communities. *BMC Bioinformatics* 9:34. DOI: 10.1186/1471-2105-9-34.
727
728 Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ. 2006.
729 Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proceedings of the*
730 *National Academy of Sciences* 103:12115–12120. DOI: 10.1073/pnas.0605127103.
731
732 Tóthmérész B. 1995. Comparison of different methods for diversity ordering. *Journal of*
733 *Vegetation Science* 6:283–290. DOI: 10.1234/12345678.
734
735 Turnbaugh PJ, Gordon JI. 2009. The core gut microbiome, energy balance and obesity. *Journal of*
736 *Physiology* 587:4153–4158. DOI: 10.1113/jphysiol.2009.174136.
737
738 Willis A, Bunge J. 2014. Estimating diversity via frequency ratios. arXiv preprint,
739 arXiv:1408.3333v2. (accessed 9 December 2014).
740

SUPPLEMENTAL INFORMATION

Supplemental Text S1. R codes for obtaining estimators of Hill numbers.

Supplemental Text S2. Simulation results based on six species abundance models.

Supplemental Table S1. Diversity analyses for the data sets in Allen et al. (2013).

PeerJ PrePrints | https://dx.doi.org/10.7287/peerj.preprints.1353v1 | CC-BY 4.0 Open Access | rec: 8 Sep 2015, publ: 8 Sep 2015