Supporting Information

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, 30043

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

To investigate the performance of the proposed singleton count estimator given in Equation (5) and the diversity estimator in Equation (7) of the main text, we carried out simulations by generating data sets from various species abundance models. Here we report the results from six representative models. In each model, we fixed the number of species at *S* = 2000 to mimic the taxa richness of microbial communities.

The functional forms or distributions for species' relative abundances $(p_1, p_2, ..., p_S)$ are given below, whereby *c* is a normalizing constant such that $\sum_{i=1}^{S} p_i = 1$. When species abundances were simulated from a distribution (Model 3 and Model 4), we first generated a set of 2000 random variables, which we regarded as fixed parameters in the simulation. In each model, we also give the CV (which is the ratio of the standard deviation over the mean) of $(p_1, p_2, ..., p_S)$. The CV value quantifies the degree of heterogeneity among the probabilities $(p_1, p_2, ..., p_S)$. When all probabilities are equal, CV = 0. A larger value of CV indicates a higher degree of heterogeneity among probabilities. In the following description, *S* = 2000 for all models.

Model 1. A homogeneous model with $p_i = 1/S$ and $S = 2000$. This is the model with no heterogeneity among species relative abundances (CV = 0).

Model 2. A random uniform model with $p_i = ca_i$, where $(p_1, p_2, ..., p_S)$ is a random sample from a uniform (0, 1) distribution. (CV = 0.57).

Model 3. A broken-stick model with $p_i = ca_i$, where $(a_1, a_2, \ldots, a_S)$ is a random sample from an exponential distribution. Equivalently, $(p_1, p_2, ..., p_S)$ follows a Dirichlet distribution with parameter 1 (CV = 0.99).

Model 4. A log-normal model with $p_i = ca_i$, where $(a_1, a_2, \ldots, a_S)$ is a random sample from a log-normal distribution with mean $\mu = 0$, and variance $\sigma^2 = 1$ (CV= 1.96).

Model 5. A Zipf-Mandelbrot model with $p_i = c/(i+5)$, $i = 1, 2, \ldots, S$ (CV = 3.07).

Model 6. A power-decay model with $p_i = c / i^{0.9}$, $i = 1, 2, \ldots, S$ (CV= 5.03).

For each given model, we considered a range of sample sizes (n = 2000 to 10000 in an increment of 1000). Then for each combination of abundance model and sample size, 1000 simulated data sets were generated from the abundance model. Two types of data were generated:

(i) True data without sequencing error (data with the true number of singletons): individuals were randomly selected from a given model; species abundances and frequency counts were then generated.

(ii) Spurious data with a sequencing error rate of 10% (data with spurious singletons): individuals were randomly selected from a given model, but there was a probability 10% that each sampled individual was misclassified as a new species and thus became a spurious singleton. This was used to mimic the sequencing error with an error rate of 10% for each detected individual to be misclassified as a spurious singleton.

For each model, we display four sub-plots in Supplementary Fig. S1: In Panel (a), we show the 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 estimation method based on Equation (5) of the main text. All values were averaged over 1000 simulation trials under the six species abundance models. All six panels (a) were also shown in Fig. 1 of the main text. Some conclusions presented there are summarized below to make the contents of this Additional file self-contained.

The number of singletons for the true data generally declines with sample size when sample size becomes very large, whereas the number of singletons for spurious data always increases with sample size. This is consistent with a similar finding by Dickie (2010). The drastically different pattern for the two singleton counts can be used to detect whether sequencing error exists or not when an empirical accumulation curve for the singleton count can be recorded in the data-collecting procedures. Fig. S1 reveals that our estimated singleton count matches very closely the true value for each model. This implies (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 observed count can be used to infer 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.

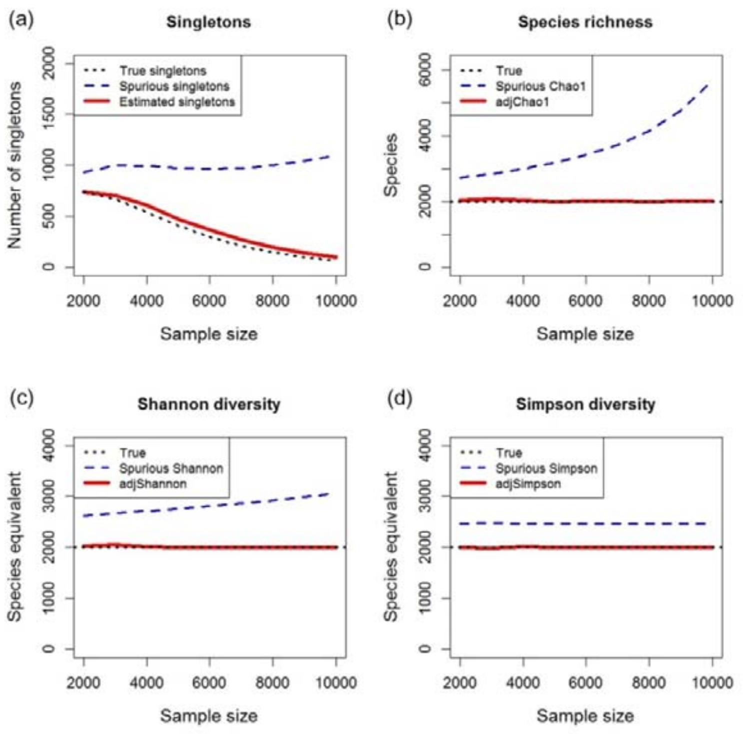
Under each model, Panels (b), (c) and (d) compare the true diversity (Equation 1 in the main text) and the estimated asymptote of diversity (Equation 7 in the main text) calculated respectively from spurious data and from the adjusted data with the observed singleton count being replaced by the estimated value computed from Equation (5) of the main text. In Panel (b), we show the plots of the true species richness and the average values (over 1000 simulation trails) of the Chao1 estimator for the spurious data and for the adjusted data. It is clear that the Chao1 estimate for the spurious data severely overestimates the true species richness. The adjusted Chao1 estimator reduces most of the positive bias and works reasonably well for all models, although negative bias exists and the magnitude of the bias increases with the CV value.

In Panel (c), we show the plots of the true Shannon diversity and the average values (over 1000 simulation trails) of the estimated asymptote of Shannon diversity for the spurious data and for the adjusted data. The estimated asymptote of Shannon estimator for spurious data moderately overestimates the true diversity for each model, but this estimated asymptote for the adjusted data exhibits very low bias and works well for all models.

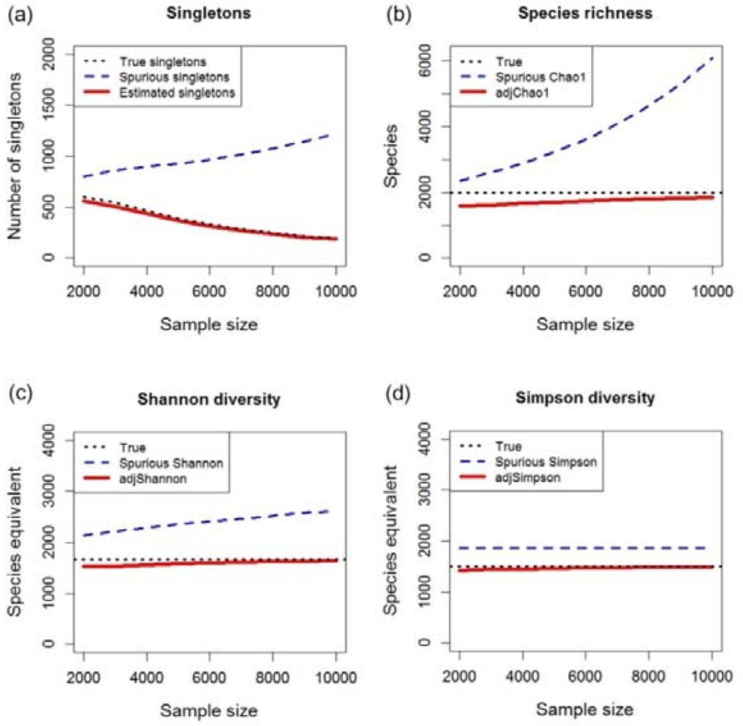
In Panel (d), we show the plots of the true Simpson diversity and the average values (over 1000 simulation trails) of the estimated asymptote of Simpson diversity for the spurious data and for the Simpson diversity estimator for the adjusted data. The estimated asymptote of Simpson estimator slightly overestimates the true diversity for each model, but this estimated asymptote for the adjusted data is nearly unbiased for all models.

In summary, our estimated asymptotes of diversities presented in Equation (7) of the main text based on the adjusted data greatly remove the positive biases due to spurious singletons. When there are sequencing errors, our procedure always leads to better results; when there are no sequencing errors, our results differ from those based on the true data only to a limited extent. Therefore, our proposed estimator of singleton count can be used to detect the quality of the observed singleton count. This also reveals that whenever singletons are uncertain or in doubt, it is worth applying our estimator of singleton count in diversity analysis and statistical inferences.

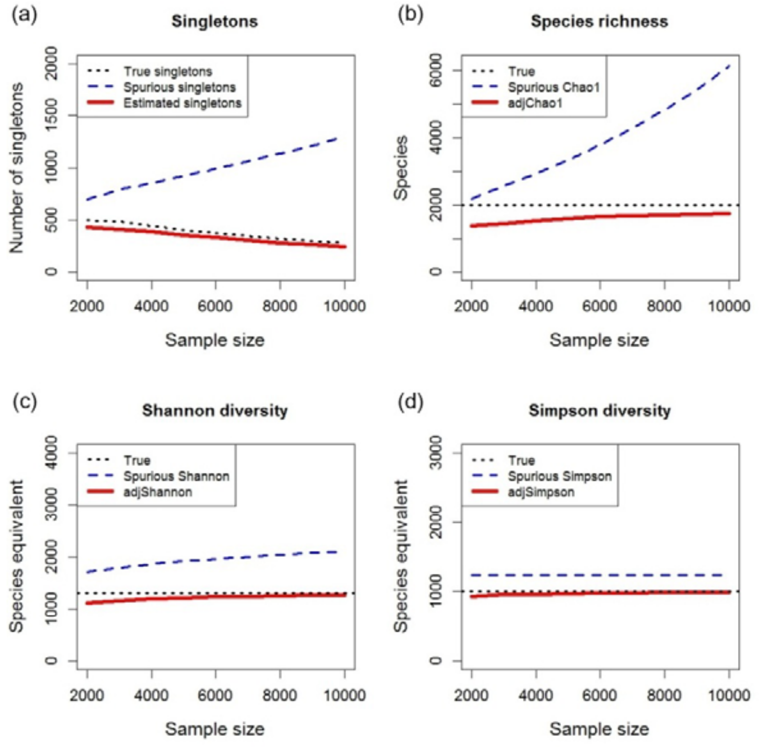
Model 1: Homogeneous model (CV=0)



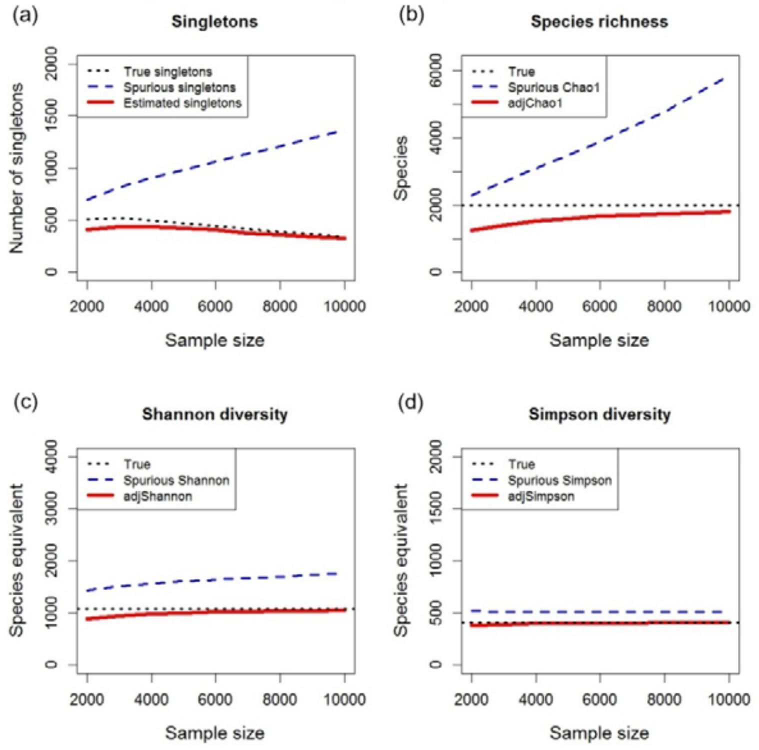
Model 2: random uniform model (CV=0.57)



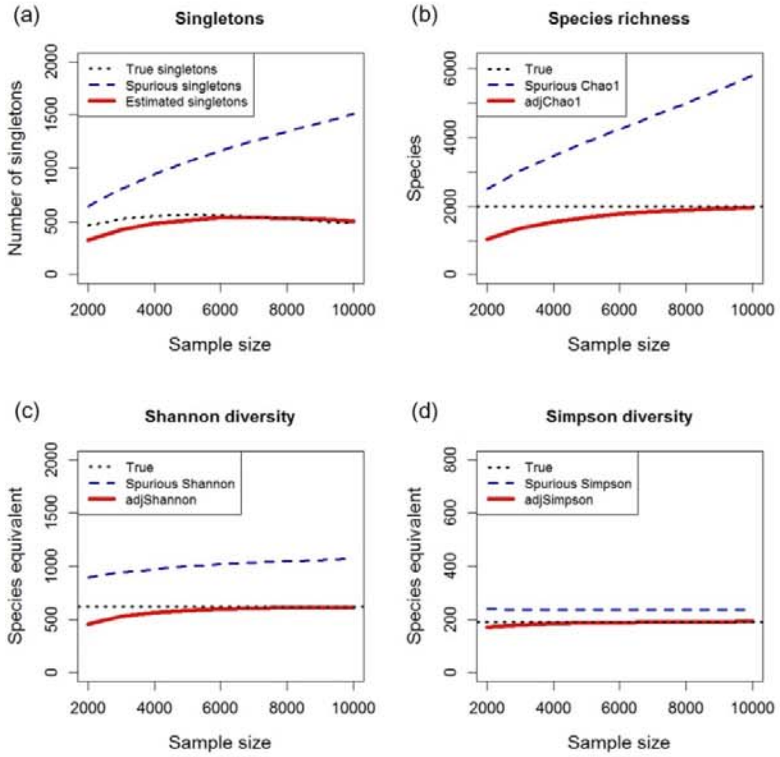
Model 3: broken-stick model (CV=0.99)



Model 4: log-normal model (CV=1.96)



Model 5: Zipf-Mandelbrot model (CV=3.07)



Model 6: power decay model (CV=5.03)

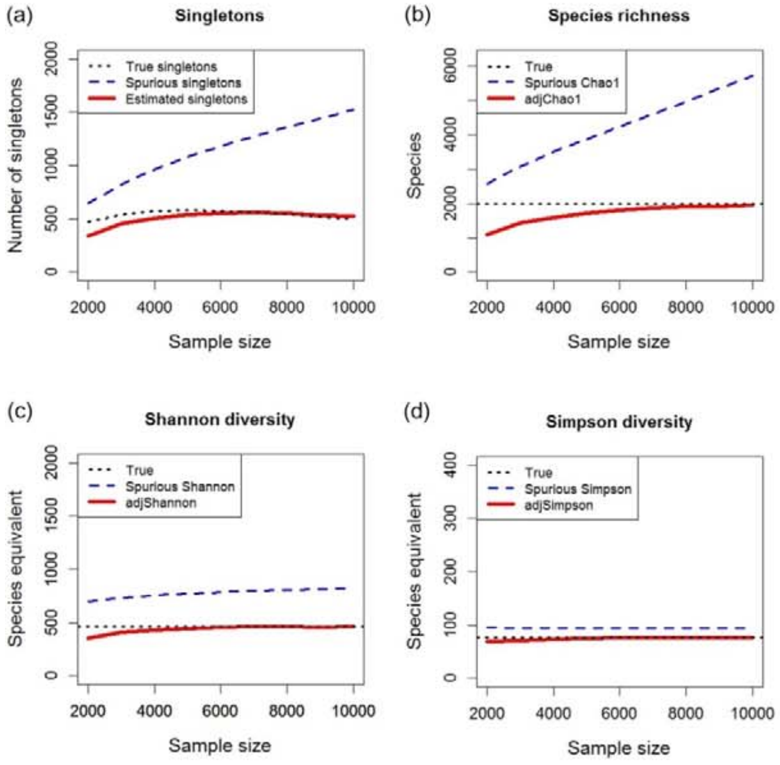


Fig S1. Plots of simulation. Under each model, there are four panels.

Panel (a): plots of the average values of the singleton counts obtained from the true data, spurious data, and the estimation method based on Equation (5) in the main text. All values represent the average values over 1000 simulation trials under six species abundance models.

Panel (b): plots of the true species richness, and the average values (over 1000 simulation trails) of the Chao1 estimator for the spurious data, and the Chao1 estimator (denoted as "adjChao1" in the plot) for the adjusted data with the observed singleton count being replaced by the estimated value computed from Equation (5) of the main text.

Panel (c): plots of the true Shannon diversity and the average values (over 1000 simulation trails) of the estimated asymptote of Shannon diversity for the spurious data, and the estimated asymptote of Shannon diversity estimator (denoted as "adjShannon" in the plot) for the adjusted data.

Panel (d): plots of the true Simpson diversity and the average values (over 1000 simulation trails) of the estimated asymptote of Simpson diversity for the spurious data, and the estimated asymptote of Simpson diversity estimator (denoted as "adjSimpson" in the plot) for the adjusted data.

Note the scale in the Y-axis may be different in the four panels due to different range of diversity.

Reference

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.