Random Sampling Process Leads to Overestimation of β-Diversity of Microbial Communities

Jizhong Zhou, Yi-Huei Jiang, Ye Deng, et al.
2013. Random Sampling Process Leads to Overestimation of β-Diversity of Microbial Communities. mBio 4(3): . doi:10.1128/mBio.00324-13.

RESEARCH ARTICLE

Random Sampling Process Leads to Overestimation of β-Diversity of Microbial Communities

Jizhong Zhou,[a,b,c] Yi-Huei Jiang,[b] Ye Deng,[b] Zhou Shi,[b] Benjamin Yamin Zhou,[b] Kai Xue,[b] Liyou Wu,[b] Zhili He,[b] Yunfeng Yang[a]

State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China[a]; Institute for Environmental Genomics and Department of Microbiology and Plant Biology, University of Oklahoma, Norman, Oklahoma, USA[b]; Earth Sciences Division, Lawrence Berkeley National Laboratory, Berkeley, California, USA[c]

ABSTRACT **The site-to-site variability in species composition, known as β-diversity, is crucial to understanding spatiotemporal patterns of species diversity and the mechanisms controlling community composition and structure. However, quantifying β-diversity in microbial ecology using sequencing-based technologies is a great challenge because of a high number of sequencing errors, bias, and poor reproducibility and quantification. Herein, based on general sampling theory, a mathematical framework is first developed for simulating the effects of random sampling processes on quantifying β-diversity when the community size is known or unknown. Also, using an analogous ball example under Poisson sampling with limited sampling efforts, the developed mathematical framework can exactly predict the low reproducibility among technically replicate samples from the same community of a certain species abundance distribution, which provides explicit evidences of random sampling processes as the main factor causing high percentages of technical variations. In addition, the predicted values under Poisson random sampling were highly consistent with the observed low percentages of operational taxonomic unit (OTU) overlap (<30% and <20% for two and three tags, respectively, based on both Jaccard and Bray-Curtis dissimilarity indexes), further supporting the hypothesis that the poor reproducibility among technical replicates is due to the artifacts associated with random sampling processes. Finally, a mathematical framework was developed for predicting sampling efforts to achieve a desired overlap among replicate samples. Our modeling simulations predict that several orders of magnitude more sequencing efforts are needed to achieve desired high technical reproducibility. These results suggest that great caution needs to be taken in quantifying and interpreting β-diversity for microbial community analysis using next-generation sequencing technologies.**

IMPORTANCE **Due to the vast diversity and uncultivated status of the majority of microorganisms, microbial detection, characterization, and quantitation are of great challenge. Although large-scale metagenome sequencing technology such as PCR-based amplicon sequencing has revolutionized the studies of microbial communities, it suffers from several inherent drawbacks, such as a high number of sequencing errors, biases, poor quantitation, and very high percentages of technical variations, which could greatly overestimate microbial biodiversity. Based on general sampling theory, this study provided the first explicit evidence to demonstrate the importance of random sampling processes in estimating microbial β-diversity, which has not been adequately recognized and addressed in microbial ecology. Since most ecological studies are involved in random sampling, the conclusions learned from this study should also be applicable to other ecological studies in general. In summary, the results presented in this study should have important implications for examining microbial biodiversity to address both basic theoretical and applied management questions.**

Received 2 May 2013 **Accepted** 7 May 2013 **Published** 11 June 2013

Citation Zhou JZ, Jiang Y, Deng Y, Shi Z, Zhou BY, Xue K, Wu L, He Z, Yang Y. 2013. Random sampling process leads to overestimation of β-diversity of microbial communities. mBio 4(3):e00324-13. doi:10.1128/mBio.00324-13.

Editor James Tiedje, Michigan State University

Address correspondence to Jizhong Zhou, jzhou@ou.edu.

Microorganisms appear to be the most diverse group of life presently known, inhabiting almost every imaginable environment on Earth (1). They play integral and unique roles in ecosystem functions and biogeochemical cycling of carbon (C), nitrogen (N), sulfur (S), phosphorus (P), and various metals. Understanding the structure, functions, interactions, stability, and adaptations of microbial populations/communities is crucial for basic science discovery (2, 3), biotechnology (4), agriculture (5), energy (6), the environment (7), and human health (8). However, due to their extremely high diversity and as-yet-uncultivated status, characterizing microbial diversity and establishing the linkages between microbial diversity and ecosystem function are very challenging (9). Understanding the mechanisms controlling microbial diversity and functions is even more difficult. The recent advances in metagenomics, which has emerged as a cutting-edge 21st century science (10), and associated metagenomics technologies such as high-throughput sequencing and microarrays (10, 11) provide revolutionary tools to address these challenges. Large-scale high-throughput sequencing-based metagenomics is providing unprecedented views of the taxonomic diversity, metabolic

potential, and ecological roles of microbial communities in various habitats (12–32). Various studies clearly demonstrate that large-scale sequencing approaches are powerful in studying microbial community diversity and activity (23, 33–43).

β-Diversity, the site-to-site variability in species composition, is crucial in understanding patterns of species diversity across various spatial and temporal scales. It can provide insight on the mechanisms controlling community compositions and structure (44). β-Diversity is widely used for investigating the mechanisms controlling biodiversity (45–47) and the responses of biological communities to environmental changes (45). In microbial ecology, high-throughput metagenomics sequencing and associated technologies are the major tools for examining the site-to-site variability in species composition and its response to environmental changes (48, 49). However, the amplicon-based approach has shown limited reproducibility, especially when examining low-abundance taxa (50). For instance, based on the overlap of operational taxonomic units (OTUs), a very low reproducibility ($<$20% between two technical replicates and $<$10% among three technical replicates) was obtained (50), which is far away from the theoretical expectation of 100% overlap among technical replicates. Similar results were recently obtained with mock communities (51) and human microbiomes (13).

The high numbers of variations in technical replicates are most likely due to the sampling artifacts associated with random sampling processes (50, 52), because many steps in the pyrotag-based sequencing analysis are associated with random sampling, e.g., PCR amplification of target genes, ligation of amplified PCR products to sequencing adaptors, emulsion and immobilization of beads, and bead deposition (50). Since microbial communities under natural settings are extremely complex and generally have abundance curves with very long tails, that is, large portions of OTUs exist in extremely low abundance, the probability of sampling such rare OTUs in a sampling event is low. The chances of resampling them are even lower, especially with limited sampling efforts (i.e., percentages of total individuals sampled in a community). It is expected that the severity of such sampling artifacts on community comparison is dependent on community complexity and sampling efforts. As the complexity of a microbial community increases, such an artifact will become more severe. Increasing sampling efforts will help to ameliorate such a problem (50). However, there is no theoretical foundation to support such a speculation.

In this study, we hypothesize that a random sampling process is the main cause for high numbers of variations among technical replicates. To test this hypothesis, the main objective of this study is to provide a theoretical foundation on understanding such sampling artifacts associated with a random sampling process. We first developed a theoretical framework to simulate the random sampling processes based on general sampling theory and to predict sampling efforts for achieving a desired reproducibility. We then illustrated the sampling artifacts associated with random sampling processes using an analogous example. We also examined whether the developed framework could be used to predict the low percentage of overlap of OTUs among technical replicates. Our results indicated that high numbers of variations in technical replicates were due to the artifacts associated with random sampling processes.

Mathematical framework. (i) Sampling individuals from a large regional community. The pattern of species abundances across different spatial-temporal scales is a central issue in ecology. However, it is generally impossible to directly measure the abundance of all species at ecologically relevant scales. Thus, it is important to understand the relationship between the underlying species abundance distribution of a large regional community and the observed abundance distribution in a small sample from the large regional community. Various statistical sampling theories are developed to describe the relationships of the species abundance between the sampled community and the large regional community (53). Here, we will use general sampling theory to simulate and predict the sampling artifacts associated with random sampling processes.

Assuming that a species (or OTU) occurs in a large regional community, the number of its individuals that exist in a small sample of the large community is dependent on the total abundance of that species in the large community, the size of the sample, and the spatial distribution of the individuals (54). In this study, we assume that all individuals in the large regional community are located randomly in space. Let N represent the total number of individuals (e.g., 16S rRNA gene sequences) and n be the number of different species (e.g., OTUs), each with abundances $x_1, x_2, \ldots, x_n$. If an individual is randomly sampled from the community, the probability of it belonging to the i^{th} species is x_i/N. If m individuals are randomly sampled from this large community, the expected number of individuals from the i^{th} species is mx_i/N. With replacement, then, the probability of the i^{th} species with abundance x_i to be encountered by k individuals in the sample is given by the following binomial function expression (55, 56):

$$p(k|x_i, N, m) = \binom{m}{k}\left(\frac{x_i}{N}\right)^k\left(1-\frac{x_i}{N}\right)^{m-k} \quad (1)$$

To estimate the probability that at least one individual of the i^{th} species is present in the sample, we generally calculate the probability that the i^{th} species is absent in the sample, i.e., $k = 0$. If $k = 0$, then the probability of the i^{th} species with abundance x_i in a community to be absent in the sample is expressed as

$$p(0|x_i, N, m) = \left(1-\frac{x_i}{N}\right)^m \quad (2)$$

Let x_i/N equal ax_i/m, where $a = m/N$, the sampling ratio. Then, equation 2 is approximated to the exponential form of the Poisson distribution:

$$\lim_{m\to\infty} p(0|x_i, N, m) = e^{-ax_i} \quad (3)$$

According to the Poisson distribution, the probability of the i^{th} species with abundance x_i in the community to have at least one individual in the sample can be expressed as follows (54):

$$\psi(a, x_i) = 1 - p(0|x_i, N, m) \approx 1 - e^{-ax_i} \quad (4)$$

The Poisson distribution is the simplest model for sampling individuals from a large regional community. Based on general sampling theory (53, 56), the abundance distribution observed in a sample that constitutes a proportion a of the large regional community can be expressed as

$$\phi_a(m) = \int_0^\infty \psi(a, x)\phi(x, \theta)dx \quad (5)$$

where $\varphi_a(m)$ is the observed species abundance distribution in a sample with m individuals sampled. $\varphi(x)$ represents the species abundance distribution with abundance x in the large regional

community, in which θ is the vector of parameters (57).

(ii) Expected species overlap among samples with the size of the large community known. Given a large community with a known underlying species abundance distribution, $\varphi(x)$, the expected proportion of species in common between samples can be theoretically predicted. In the case in which the species abundances in these samples are perfectly correlated, the expected number of the species shared between two samples is given by (54)

$$\int_0^\infty \psi(a_1, x)\psi(a_2, x)\phi(x, \theta)dx \tag{6}$$

and the number of species shared among three samples is

$$\int_0^\infty \psi(a_1, x)\psi(a_2, x)\psi(a_3, x)\phi(x, \theta)dx \tag{7}$$

where a_1, a_2, and a_3 are sample ratios from samples 1, 2 and 3, and $a_1 = m_1/N, l = 1, 2, 3$. m_1 is the number of individuals (i.e., 16S rRNA gene sequences in this study) obtained from the l^{th} sample.

Various similarity metrics are used for assessing overlap among different samples (58). In this study, the two popular similarity metrics, Jaccard's incidence-based and Bray-Curtis's abundance-based methods, are used. Based on the Jaccard similarity index, the proportion of species (i.e., OTUs in this study) overlap between two samples $[O_J^2(a_1,a_2,\theta)]$ is calculated as follows:

$$O_J^2(a_1, a_2, \theta) = \frac{\displaystyle\int_0^\infty \psi(a_1, x)\psi(a_2, x)\phi(x, \theta)dx}{\displaystyle\int_0^\infty \psi(a_1, x)\phi(x, \theta)dx + \int_0^\infty \psi(a_2, x)\phi(x, \theta)dx - \int_0^\infty \psi(a_1, x)\psi(a_2, x)\phi(x, \theta)dx} \tag{8}$$

The proportion of species overlap among three samples $[O_J^3(a_1,a_2,a_3,\theta)]$ is

$$O_J^3(a_1, a_2, a_3, \theta) = \frac{\displaystyle\int_0^\infty \psi(a_1, x)\psi(a_2, x)\psi(a_3, x)\phi(x, \theta)dx}{D_3} \tag{9}$$

where

$$D_3 = \int_0^\infty \psi(a_1, x)\phi(x, \theta)dx + \int_0^\infty \psi(a_2, x)\phi(x, \theta)dx + \int_0^\infty \psi(a_3, x)\phi(x, \theta)dx - \int_0^\infty \psi(a_1, x)\psi(a_2, x)\phi(x, \theta)dx - \int_0^\infty \psi(a_2, x)\psi(a_3, x)\phi(x, \theta)dx - \int_0^\infty \psi(a_1, x)\psi(a_3, x)\phi(x, \theta)dx + \int_0^\infty \psi(a_1, x)\psi(a_2, x)\psi(a_3, x)\phi(x, \theta)dx$$

There are seven commonly used continuous species abundance distributions, including lognormal, exponential, gamma, truncated hyperbolic, continuous log series (54), inverse gamma, and inverse Gaussian distribution. When the total number of individuals of the regional community (N) is known, the Jaccard similarity-based explicit expression functions of the proportion of species overlap between two or three samples are summarized in Tables S1A and S1B in the supplemental material.

(iii) Expected species overlap among samples with the size of the large community unknown. Using equations 8 and 9 to estimate the percentages of species overlap between two or three samples requires information about the total individuals (N) of the communities examined. However, in most cases, the number of total individuals in a community is unknown. In the following, we will consider the situation in which N is unknown.

Since most of the species abundance distributions are scale invariant under a Poisson sampling process (53, 57, 59, 60), the expected sample abundance distribution can be obtained by rescaling the community abundance distribution (x) to the sample abundance distribution (y): $y = px$, where p is the proportion of the community sampled. Thus, equation 8 can be rewritten as

$$O_J^{2*}(a_1^*, a_2^*, \theta^*) = \frac{\displaystyle\int_0^\infty \psi(a_1^*, y)\psi(a_2^*, y)\phi(y, \theta^*)dy}{\displaystyle\int_0^\infty \psi(a_1^*, y)\phi(y, \theta^*)dy + \int_0^\infty \psi(a_2^*, y)\phi(y, \theta^*)dy - \int_0^\infty \psi(a_1^*, y)\psi(a_2^*, y)\phi(y, \theta^*)dy} \tag{10}$$

where a_1^*, a_2^*, and θ^* are the sample ratios and the vector of parameters when N is unknown, which are the rescaled parameters and functions of p. In the case of two random samples from the same community with parameters θ, we set $p = a_1 + a_2$ and then obtain $a_1^* = a_1/(a_1 + a_2)$ and $a_2^* = a_2/(a_1 + a_2)$. Let m_1 and m_2 be the total number of individuals observed in sample 1 and sample 2, respectively. Now we can substitute $a_1 = m_1/N$ and $a_2 = m_2/N$ into and a_2^*, and then both a_1^* and a_2^* are the functions of m_1 and m_2, i.e., $a_1^* = m_1/(m_1 + m_2)$ and $a_2^* = m_2/(m_1 + m_2)$. θ^* is the vector of parameters of the species abundance distribution in the samples. Consequently, the expected species overlap can be obtained by fitting data to the parameters θ^* without knowing the total number of individuals of the community, N.

Similarly, in the case of three samples randomly sampled from one community, the parameter p can be set to the sum of sampling ratios, $p = a_1 + a_2 + a_3$. Then, equation 9 can be rewritten as

$$O_J^{3*}(a_1^*, a_2^*, a_3^*, \theta^*) = \frac{\displaystyle\int_0^\infty \psi(a_1^*, y)\psi(a_2^*, y)\psi(a_3^*, y)\phi(y, \theta^*)dy}{d_3} \tag{11}$$

where

$$
\begin{aligned}
d_3 = & \int_0^{\infty} \psi(a_1^*, y)\phi(y, \theta^*)dy \\
& + \int_0^{\infty} \psi(a_2^*, y)\phi(y, \theta^*)dy \\
& + \int_0^{\infty} \psi(a_3^*, y)\phi(y, \theta^*)dy \\
& - \int_0^{\infty} \psi(a_1^*, y)\psi(a_2^*, y)\phi(y, \theta^*)dy \\
& - \int_0^{\infty} (a_1^*, y)\psi(a_3^*, y)\phi(y, \theta^*)dy \\
& - \int_0^{\infty} \psi(a_2^*, y)\psi(a_3^*, y)\phi(y, \theta^*)dy \\
& + \int_0^{\infty} \psi(a_1^*, y)\psi(a_2^*, y)\psi(a_3^*, y)\phi(y, \theta^*)dy
\end{aligned} \tag{12}
$$

and $a_1^* = a_1/(a_1 + a_2 + a_3) = m_1/(m_1 + m_2 + m_3)$, $a_2^* = a_2/(a_1 + a_2 + a_3) = m_2/(m_1 + m_2 + m_3)$, $a_3^* = a_3/(a_1 + a_{2+} + a_3) = m_3/(m_1 + m_2 + m_3)$.

Now, according to equations 10 and 11, the species overlap for two and three samples can be estimated based on sample abundance y rather than community abundance x, so that there is no need to know the community size N. When the total number of individuals of the regional community is unknown, the Jaccard similarity-based explicit expression functions of the proportion of species overlap between two or among three samples are summarized in Tables S1C and S1D in the supplemental material.

(iv) Predicting sampling efforts for achieving a desired overlap among replicate samples. One important question in practice is what levels of sampling efforts are needed for achieving desired species overlap when the number of total individuals of the community is unknown. In the following, we will address this practical question by assuming that and m_1' and m_2' individuals are needed for sampling to achieve a desired overlap between two samples. To simplify the situation, let $m_1' = m_2' = m'$, and sampling ratio $A = m'/N$. Based on equation 13, when N is known, the predicted overlap between two samples is given by

$$
O_J^{2\,'} = \frac{\int_0^{\infty} \psi(A, x)\psi(A, x)\phi(x, \theta)dx}{\int_0^{\infty} \psi(A, x)\phi(x, \theta)dx + \int_0^{\infty} \psi(A, x)\phi(x, \theta)dx - \int_0^{\infty} \psi(A, x)\psi(A, x)\phi(x, \theta)dx} \tag{13}
$$

Similarly, when N is unknown, we can take the transformation of the abundance variable $y = px = (a_1 + a_2)x$. By comparing equations 8 and 10, the relationship of the expected number of the shared species when N is known or unknown is given by

$$
\int_0^{\infty} \psi(A, x)\phi(x, \theta)dx = \int_0^{\infty} \psi(A^*, y)\phi(y, \theta^*)dy \tag{14}
$$

where $A^* = \frac{A}{a_1 + a_2} = m'/(m_1 + m_2)$ is the predicted sample ratio for achieving the desired overlap between two samples, and θ^* is the vector of scaled parameters. Rearranging A^* as the expression of the number of sequences, $A^* = m'/(m_1 + m_2)$, the predicted overlap model is given by

$$
O_J^{2\,'*} = \frac{\int_0^{\infty} \{\psi[m'/(m_1 + m_2), y]\}^2\phi(y, \theta^*)dy}{2\int_0^{\infty} \{\psi[m'/(m_1 + m_2), y]\}\phi(y, \theta^*)dy - \int_0^{\infty} \{\psi[m'/(m_1 + m_2), x]\}^2\phi(y, \theta^*)dy} \tag{15}
$$

Similarly, the predicted overlap model for three samples is given by

$$
\begin{aligned}
O_J^{3\,'*} = & \frac{\int_0^{\infty} \{\psi[m'/(m_1 + m_2 + m_3), y]\}^3\phi(y, \theta^*)dy}{3\int_0^{\infty} \psi[m'/(m_1 + m_2 + m_3), y]\phi(y, \theta^*)dy} \\
& - 3\int_0^{\infty} \{\psi[m'/(m_1 + m_2 + m_3), y]\}^2\phi(y, \theta^*)dy \\
& + \int_0^{\infty} \{\psi[m'/(m_1 + m_2 + m_3), y]\}^3\phi(y, \theta^*)dy
\end{aligned} \tag{16}
$$

By solving equations 15 and 16, the sampling efforts m' can be estimated based on sample abundance y without knowing the community size N. The Jaccard similarity-based explicit expression functions of the proportion of species overlap between two or among three samples for predicting sampling efforts are summarized in Tables S1C and S1D in the supplemental material.

RESULTS

Simulation with an analogous example. To better illustrate the effects of random sampling processes on the OTU overlap among technical replicates (3), we use an analogous example by randomly sampling balls from three jars containing the same number of balls of different colors (Fig. 1). Assume that three identical jars contain N balls of n different colors. Their abundance distributions vary among different colors of balls but are identical among these jars. Here, individual balls are equivalent to individual 16S rRNA gene sequences, while the types of balls with different colors are equivalent to individual OTUs. To simplify the situation, we assume that the same numbers of balls, m, are randomly sampled from these three jars, yielding samples 1, 2, and 3 (Fig. 1). Theoretically, if all balls from the whole jars are sampled (i.e., $m = N$), 100% ball overlap will be expected among these samples. However, in reality, the percentages of ball overlap will be less than 100%, because they depend on the sampling efforts, ball abundance distribution, and

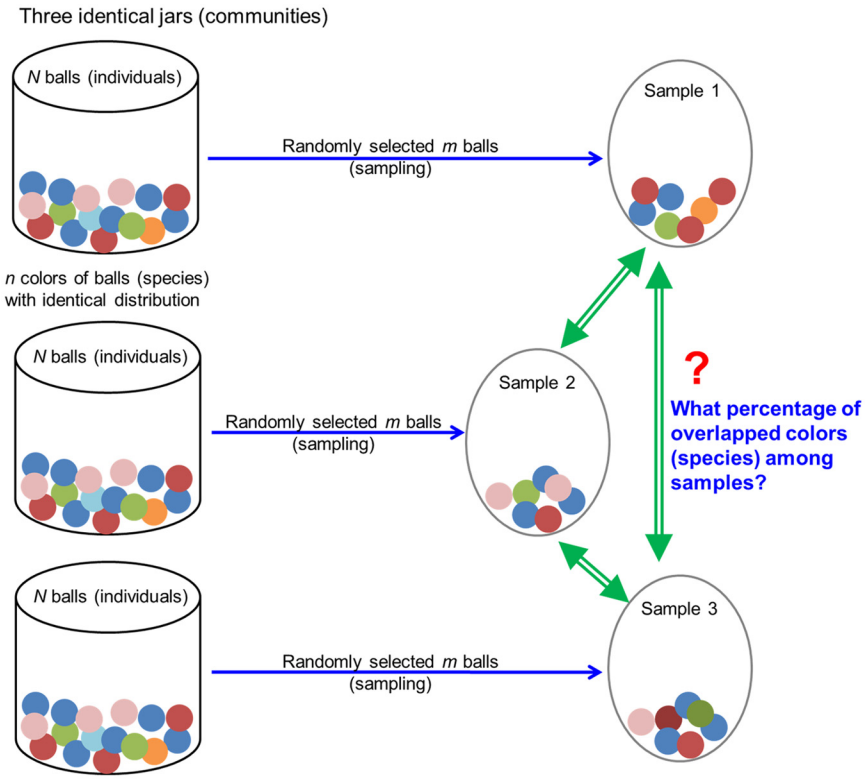

FIG 1 An analogous example to simulate random sampling processes. Three identical jars contain the same number and types of balls, with identical ball abundance distribution.

complexity of the community. The differences of the percentages of ball overlap between the theoretically predicted and the observed values among different sampling events are entirely due to random sampling processes, because there is no difference in the ball compositions and abundances among these three jars.

With the assumption that the ball abundance distributions in these three jars follow any of the five continuous species abundance distributions as listed in Table S1A in the supplemental material, we simulated the effects of the random sampling processes on the ball overlap based on Jaccard and Bray-Curtis indexes for five different distributions: exponential, gamma, lognormal, inverse Gaussian, and inverse gamma. Under each specific ball abundance distribution, the average observed overlap through simulations of 100 repeated samplings (see Materials and Methods for details) was calculated and compared to the theoretically predicted overlap between two samples based on equation 8 or 10, when N is known or unknown, respectively. Similar analyses were carried out for comparing ball overlaps among three samples based on equation 9 or 11. Although the simulation results vary considerably with the parameters selected, the following generalizations can be drawn. First, no significant differences of Jaccard similarities were observed between the theoretically predicted and the observed overlap values for all five species abundance distributions (Table 1 and Fig. 2; see also Fig. S1 and S2 in the supplemental material), indicating that the theoretically predicted percentages of ball overlap match well to the observed percentages of ball overlap for all ball abundance distributions examined. Second, there were no significant differences of the predicted overlap percentages derived from equation 8 with known N and equation 10 with unknown N for two samples or from equation 9 with known N and equation 11 with unknown N for three samples (see Table S2 in the supplemental material), suggesting that accurate predictions of the overlap percentages between two or among three samples can be obtained when the size of the regional communities is unknown. In addition, low ball overlap percentages were observed with low sampling efforts under different ball abundance distributions (Fig. 2; see also Fig. S1 and S2). For instance, under exponential distribution, when 1% of the community was sampled, 50% overlap between two samples (Fig. 2A) and 34% overlap among three samples (Fig. 2B) were obtained. As the sampling efforts increase, the ball overlap among samples increases under different ball abundance distributions (Fig. 2; see also Fig. S1 and S2). For example, under exponential distribution, when 10% of the community was sampled, 91% ball overlap between two samples (Fig. 2A) and 82% among three samples were obtained (Fig. 2B). All of the above results suggested that accurate predictions of the overlap could be obtained between two samples with equation 8 or 10 and among three samples with equation 9 or 11.

Because the general explicit form of the ball overlap for quantitative similarity index could not be derived, numerical simulations were performed (see Materials and Methods) for all five ball abundance distributions (see Fig. S3 in the supplemental material)

TABLE 1 Chi-square-based goodness-of-fit test of the observed and predicted percentages of overlap for the analogous example[a]

OTU abundance distribution	Two samples				Three samples			
	Known N		Unknown N		Known N		Unknown N	
	χ^2	P	χ^2	P	χ^2	P	χ^2	P
Exponential	6.6×10^{-5}	0.999	6.6×10^{-5}	0.999	2.7×10^{-4}	0.999	2.7×10^{-4}	0.999
Gamma	1.9×10^{-6}	0.999	5.1×10^{-4}	0.999	9.4×10^{-6}	0.999	6.5×10^{-3}	0.999
Lognormal	4.2×10^{-4}	0.999	2.1×10^{-4}	0.999	1.1×10^{-3}	0.999	1.1×10^{-3}	0.999
Inverse gamma	4.0×10^{-6}	0.999	2.1×10^{-4}	0.999	6.5×10^{-6}	0.999	2.2×10^{-3}	0.999
Inverse Gaussian	3.3×10^{-5}	0.999	9.1×10^{-4}	0.999	7.5×10^{-5}	0.999	8.5×10^{-4}	0.999

[a] Detailed information is presented in Fig. S2 and S3 in the supplemental material.

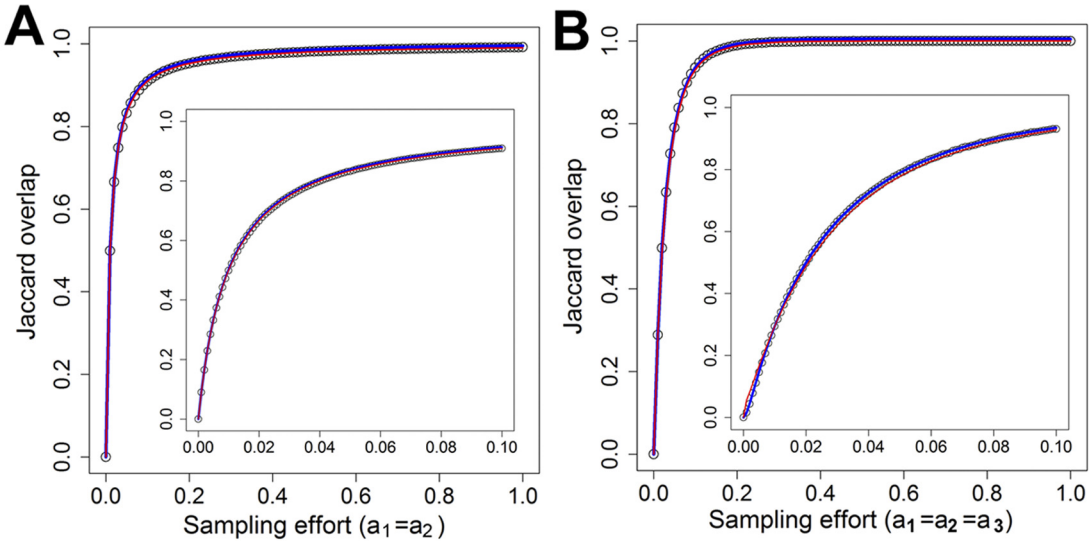

FIG 2 The relationships between the expected Jaccard overlaps of ball colors and sampling efforts under the exponential abundance distribution, assuming the community has 10^6 individual balls and 10^4 types of balls, with different colors. Distribution parameter is set to $\lambda = 1 \times 10^{-2}$. In each case, we calculated the theoretically predicted overlap (blue line) by equation 8 when N is known, the predicted overlap (red line) by equation 10 when N is unknown, and the average observed overlap (point) through simulations of 100 repeated samplings. (A) Two samples. The sample ratio is $a_1 = a_2$. (B) Three samples. The sample ratio is $a_1 = a_2 = a_3$.

for comparisons based on Jaccard similarity, all of the numerical simulations indicated that low percentages of overlap were also observed among samples with low sampling efforts. For instance, under exponential distribution, when 1% of the community was sampled, 53% of ball overlap based on the Bray-Curtis index was obtained between two samples and 35% among three samples. All of these results, based on both Jaccard and Bray-Curtis similarities, indicate that substantially high numbers of variations can be obtained among communities of identical compositions and abundance distributions when sampling effort is low. Such variation is entirely due to random sampling processes.

Empirical examples. In our previous studies, the composition and structure of 24 microbial communities from a long-term global change experimental site in Oklahoma were analyzed using the amplicon-based sequencing detection approach (49). Each community was amplified with 2 or 3 bar-coded primers, followed by sequencing using both forward and reverse primers. Thus, three types of datasets were available: sequences from forward and reverse primers and combined sequences. Since the sequences of individual technical replicates within an experimental plot are derived from the same community, conceptually, they should obey the same species abundance distribution. Thus, the sequences from all technical replicates within a plot (a community) were pooled for fitting species abundance models described in Table S1A in the supplemental material. The best-fit model and associated parameters for each plot (i.e., community) were shown in Table S3A for the soil communities amplified with two tags and in Table S3B for the soil communities amplified with three tags. Either the exponential abundance distribution or the inverse gamma distribution is the preferred OTU abundance distribution for all soil samples (see Tables S3A and S3B).

Once the OTU abundance distribution models are determined for all communities examined, the predicted overlap percentages in terms of Jaccard and Bray-Curtis similarities were calculated based on the formula provided in Tables S1C and S1D in the supplemental material. The results were listed in Tables S4A and S4B for the communities amplified with two tags and in Tables S5A and S5B for those with three tags. Overall, no significant differences were observed between observed and predicted Jaccard overlaps (Table 2).

Since Bray-Curtis overlap is a quantitative similarity index, it is generally higher than the incidence-based overlap (see Tables S5A and S5B in the supplemental material). Overall, no significant differences were observed between observed and predicted Bray-Curtis overlap percentages (Table 2). All of the above-reported results indicated that the predicted overlap percentages match well to the observed results. Therefore, the variations among tech-

TABLE 2 Chi-square-based goodness-of-fit test of the observed and predicted percentages of overlap for the experimental data[a]

Communities amplified with:	Similarity test	Forward primer		Reverse primer		Combined	
		χ^2	P	χ^2	P	χ^2	P
Two tags	Jaccard	0.040	0.999	0.060	0.999	0.038	0.999
	Bray-Curtis	0.106	0.999	0.122	0.999	0.075	0.999
Three tags	Jaccard	0.006	0.999	0.010	0.999	0.005	0.999
	Bray-Curtis	0.008	0.999	0.026	0.999	0.016	0.999

[a] Detailed data are listed in Table S4A (two tags, Jaccard), S4B (two tags, Bray-Curtis), S5A (three tags, Jaccard), and S5B (three tags, Bray-Curtis) in the supplemental material.

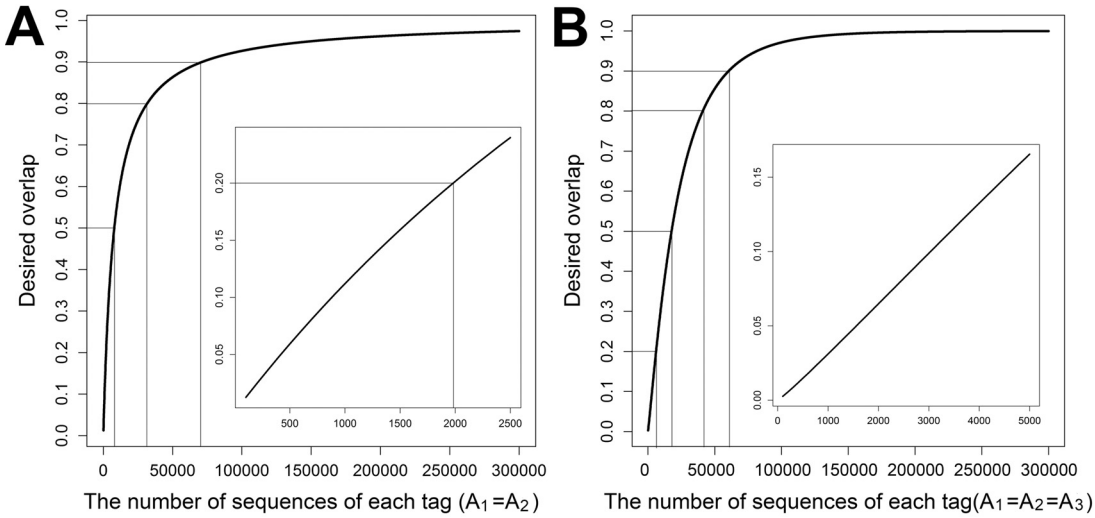

FIG 3 Prediction of sampling efforts for desired OTU overlap. (A) Desired overlap between two tags based on the combined sequences from sample 2UC. The sampling efforts were calculated based on equation 15. The parameters for species abundance distribution were from Table S3A in the supplemental material. (B) Desired overlap among three tags based on the combined sequences from sample 1UC. The sampling efforts were calculated based on equation 16. The parameters for species abundance distribution were from Table S3B.

nical replicates can be best explained by the artifacts associated with random sampling processes.

Predictions of sampling efforts for the desired species overlap. Based on equations 15 and 16, on the explicit expression functions presented in Tables S1C and S1D, and on sample abundance and sequence information for various tags in Tables S3A and S3B, sampling efforts for achieving various degrees of OTU overlaps between two (see Table S6A) and among three technical replicates (see Table S6B) were predicted. To achieve 20% OTU overlap for two tags, an average of 2,367 sequences are needed (see Table S6A; Fig. 2A, 2UC), which is consistent with the number of sequences obtained in this experiment. To achieve 90% overlap between two technical replicates, an average of 60,900 sequences are needed (see Table S6A). For instance, for the community of 2UC, a total of 71,400 of the combined sequences would be needed to achieve 90% OTU overlap between two technical replicates (Fig. 3A), which is about 32 times more sequence reads needed than we sequenced previously (see Table S4 to S6).

Much more sequencing effort is needed to achieve desired OTU overlaps among three technical replicates than between two technical replicates. To have 10% OTU overlap for three tags, an average of 3,310 sequences are needed, which is consistent with the number of sequences obtained in this experiment. To reach 90% overlap between three technical replicates, an average of 63,770 sequences are needed (see Table S6B). For example, for the community of 1UC, about 60,500 sequences are required to obtain 90% of OTU overlap (Fig. 3B). The current sampling efforts (2,018 sequences) are far less than the desired 90% OTU overlap. Our results also suggested that most of the work published, especially with soils, is severely undersampling if the goal is to determine significant changes in β-diversity among sampling sites.

DISCUSSION

One of the major technical challenges for the amplicon-based sequencing detection approach is low reproducibility (50), which is a central issue in comparative studies (61). This issue has recently been examined intensively, but it is still a matter of debate (13, 31, 51). Results from several recent studies supported (13, 50, 51, 62), disputed (31, 63–65) or both supported and disputed (66) our previous observations. A number of factors can contribute to such divergent observations, e.g., the complexity of the systems examined (50, 51, 64), differences in sequencing depths (50, 62, 63), and/or variations in sequencing and sequence preprocessing approaches (51, 67). Since most of these factors act with each other, isolating individual factors influencing sequencing reproducibility is extremely difficult, especially when natural communities of unknown diversity background are examined. Using artificial communities of known diversity, Pinto and Raskin (51) provided explicit evidences for the poor reproducibility of the amplicon sequencing-based detection approach, even with the simple community and relatively deep sequencing. Thus, poor reproducibility is a problem inherent in the amplicon sequencing-based detection approach (50).

Such poor reproducibility among technical replicates could result from PCR amplification biases (67–72), sequencing errors (51, 67, 71, 72), and/or the artifacts associated with random sampling processes during sample preparation and sequencing (50, 52). In this study, using analogous ball examples, we showed that very low percentages of overlap were observed among replicate samples from the community with identical ball types and numbers when the sampling effort was low, which is very consistent with what we observed experimentally (50). In addition, under different OTU abundance distributions, the obtained overlap percentages among two or three technical replicates were very consistent with the theoretical predicted values under the assumption of random sampling processes. The simulation results presented in this study provided explicit evidences of the contributions of random sampling processes to the high numbers of variations observed among technical replicate samples. It should be noted that spurious OTUs due to sequencing errors could also contribute to such technical variations, but they will be indistinguishable

from rare abundant OTUs which can be detected only sporadically in replicate samples (51).

The low reproducibility of the amplicon sequencing-based detection approach associated with random sampling processes raises a concern of comparing the β-diversity of microbial communities across different samples for amplicon sequencing. Although the inherent high numbers of variations associated with random sampling processes have less effect on α-diversity, it is problematic in estimating β-diversity (50, 51), which presents a significant challenge for comparative studies across different spatial and temporal scales. However, the degrees of such effects on estimating β-diversity are dependent on the complexity of the community examined and sampling efforts. In general, as the complexity increases, such problems will be more severe (50) and greater sampling efforts will be needed (50, 62). Nevertheless, great caution needs to be taken for estimating community based on OTUs when the sequences examined are not enough to represent community diversity (51, 62).

Determining the patterns and distribution of species abundance is a central issue in ecology (53), because they are important in studying both basic ecological theory and applied biodiversity conservation. However, direct measurement of species abundance in ecologically relevant scales is difficult, if not impossible. Instead, the distribution of species abundance in ecology is generally inferred based on limited samples under Poisson sampling. Therefore, the artifacts associated with random sampling processes observed in this study for microbial communities should also be applicable to the other ecological studies in plants and animals, although such problems could be less severe in macroecology. However, to the best of our knowledge, such an issue has not been addressed in the ecological literature, but it is of critical importance in studying species distribution, especially species-area relationships (52, 73). Along with our previous efforts (50, 52), this study clearly demonstrates the importance of random sampling processes in estimating microbial biodiversity, especially β-diversity. The general conclusions learned from this study should be applicable to other ecological studies in general.

In conclusion, the factors causing variations in β-diversity are among the most important but poorly understood issues in ecology, because they are the key mechanisms influencing global variation in biodiversity. Next-generation sequencing technologies, such as PCR amplicon sequencing-based detection approaches, have been rapidly used for characterizing microbial biodiversity, but they also suffer from several inherent drawbacks, such as a high number of sequencing errors, biases, and poor reproducibility and quantitation. Through mathematical modeling and simulation based on general sampling theory, this study provides explicit evidences of random sampling processes as the main factor causing high percentages of technical variations and develops a framework for predicting sampling efforts for achieving the desired technical reproducibility. Since most ecological studies are involved in random sampling, the artifacts associated with random sampling processes observed in microbial ecology should also be applicable to macroecology, although such problems could be less severe there. Because such artifacts greatly overestimate β-diversity, great caution should be taken when the amplicon sequencing-based detection is used for drawing quantitative conclusions about β-diversity. Increasing sampling efforts and/or the number of sample replicates (both technical and biological) should be the most effective ways to ameliorate technical reproducibility for drawing more reliable quantitative conclusions, but how to balance the sampling efforts and number of samples analyzed per sequencing run is dependent on biological questions and objectives, as well as the complexity and similarity of communities examined.

MATERIALS AND METHODS

The details for all materials and methods used in this study are provided in the supplemental material (see Text S1). Briefly, based on general sampling theory (54–56), a mathematical framework was first developed under 7 different OTU abundance distributions for simulating the effects of random sampling processes on quantifying β-diversity when the community size is known or unknown. Second, an analogous ball example was used to explicitly illustrate the effects of random sampling processes on the OTU overlap among technical replicates. Third, the theoretical models were fitted with the empirical experimental sequencing data from our previous studies (50), in which a total of 24 soil communities from a long-term climate change experiment facility (74) were sequenced with 60 tags. An average of 1,121 $\pm$ 390 OTUs were obtained for each tag based on the combined samples. In addition, a χ^2 test was employed to determine whether the predicted OTU overlaps were consistent with the observed values.

SUPPLEMENTAL MATERIAL

Supplemental material for this article may be found at http://mbio.asm.org/lookup/suppl/doi:10.1128/mBio.00324-13/-/DCSupplemental.

Text S1, DOCX file, 0.1 MB.
Figure S1, DOCX file, 0.4 MB.
Figure S2, DOCX file, 0.4 MB.
Figure S3, DOCX file, 0.9 MB.
Table S1, DOCX file, 0.1 MB.
Table S2, DOCX file, 0.1 MB.
Table S3, DOCX file, 0.1 MB.
Table S4, DOCX file, 0.1 MB.
Table S5, DOCX file, 0.1 MB.
Table S6, DOCX file, 0.1 MB.

ACKNOWLEDGMENTS

This work has been supported through contract DE-AC02-05CH11231 (as part of ENIGMA, a Scientific Focus Area) and contract DE-SC0004601, by the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research, Genomics: GTL Foundational Science, by the Biological Systems Research on the Role of Microbial Communities in Carbon Cycling Program (DE-SC0004601), by the U.S. National Science Foundation under contract NSF EF-1065844, and by the State Key Joint Laboratory of Environment Simulation and Pollution Control (grant 11Z03ESPCT) at Tsinghua University.

REFERENCES

1. **Whitman WB, Coleman DC, Wiebe WJ.** 1998. Prokaryotes: the unseen majority. Proc. Natl. Acad. Sci. U. S. A. **95:**6578–6583.
2. **Zhou J, Deng Y, Luo F, He Z, Tu Q, Zhi X.** 2010. Functional molecular ecological networks. mBio **1**(4):e00169-10. http://dx.doi.org/10.1128/mBio.00169-10.
3. **Zhou J, Deng Y, Luo F, He Z, Yang Y.** 2011. Phylogenetic molecular ecological network of soil microbial communities in response to elevated CO_2. mBio **2**(4):e00122-11. http://dx.doi.org/10.1128/mBio.00122-11.
4. **Curtis TP, Head IM, Graham DW.** 2003. Theoretical ecology for engineering biology. Environ. Sci. Technol. **37:**64A–70A.
5. **Kennedy AC, Smith KL.** 1995. Soil microbial diversity and the sustainability of agricultural soils. Plant Soil **170:**75–86.
6. **Werner JJ, Knights D, Garcia ML, Scalfone NB, Smith S, Yarasheski K, Cummings TA, Beers AR, Knight R, Angenent LT.** 2011. Bacterial community structures are unique and resilient in full-scale bioenergy systems. Proc. Natl. Acad. Sci. U. S. A. **108:**4158–4163.
7. **Xu M, Wu WM, Wu L, He Z, Van Nostrand JD, Deng Y, Luo J, Carley J, Ginder-Vogel M, Gentry TJ, Gu B, Watson D, Jardine PM, Marsh TL,**

Tiedje JM, Hazen T, Criddle CS, Zhou J. 2010. Responses of microbial community functional structures to pilot-scale uranium in situ bioremediation. ISME J. **4:**1060–1070.
8. **Ley RE, Turnbaugh PJ, Klein S, Gordon JI.** 2006. Microbial ecology: human gut microbes associated with obesity. Nature **444:**1022–1023.
9. **Levin SA.** 2006. Fundamental questions in biology. PLoS Biol. **4:**e300. http://dx.doi.org/10.1371/journal.pbio.0040300.
10. **Handelsman J.** 2007. The new science of metagenomics: revealing the secrets of our microbial planet. U.S. National Academies, Washington, DC.
11. **He Z, Gentry TJ, Schadt CW, Wu L, Liebich J, Chong SC, Huang Z, Wu W, Gu B, Jardine P, Criddle C, Zhou J.** 2007. GeoChip: a comprehensive microarray for investigating biogeochemical, ecological and environmental processes. ISME J. **1:**67–77.
12. **Andersson AF, Lindberg M, Jakobsson H, Bäckhed F, Nyrén P, Engstrand L.** 2008. Comparative analysis of human gut microbiota by barcoded pyrosequencing. PLoS One **3:**e2836. http://dx.doi.org/10.1371/journal.pone.0002836.
13. **Flores GE, Henley JB, Fierer N.** 2012. A direct PCR approach to accelerate analyses of human-associated microbial communities. PLoS One **7:**e44563. http://dx.doi.org/10.1371/journal.pone.0044563.
14. **Turnbaugh PJ, Gordon JI.** 2008. An invitation to the marriage of metagenomics and metabolomics. Cell **134:**708–713.
15. **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:**480–484.
16. **Turnbaugh PJ, Ley RE, Hamady M, Fraser-Liggett CM, Knight R, Gordon JI.** 2007. The human microbiome project. Nature **449:**804–810.
17. **Arumugam M, et al.** 2011. Enterotypes of the human gut microbiome. Nature **473:**174–180.
18. **Belda-Ferre P, Alcaraz LD, Cabrera-Rubio R, Romero H, Simón-Soro A, Pignatelli M, Mira A.** 2012. The oral metagenome in health and disease. ISME J. **6:**46–56.
19. **Greenblum S, Turnbaugh PJ, Borenstein E.** 2012. Metagenomic systems biology of the human gut microbiome reveals topological shifts associated with obesity and inflammatory bowel disease. Proc. Natl. Acad. Sci. U. S. A. **109:**594–599.
20. **Gevers D, Cohan FM, Lawrence JG, Spratt BG, Coenye T, Feil EJ, Stackebrandt E, Van de Peer Y, Vandamme P, Thompson FL, Swings J.** 2005. Opinion: re-evaluating prokaryotic species. Nat. Rev. Microbiol. **3:**733–739.
21. **Koren O, Spor A, Felin J, Fåk F, Stombaugh J, Tremaroli V, Behre CJ, Knight R, Fagerberg B, Ley RE, Bäckhed F.** 2011. Human oral, gut, and plaque microbiota in patients with atherosclerosis. Proc. Natl. Acad. Sci. U. S. A. **108:**4592–4598.
22. **Thompson JR, Pacocha S, Pharino C, Klepac-Ceraj V, Hunt DE, Benoit J, Sarma-Rupavtarm R, Distel DL, Polz MF.** 2005. Genotypic diversity within a natural coastal bacterioplankton population. Science **307:**1311–1313.
23. **Qin JJ, Li RQ, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li JH, Xu JM, Li SC, Li DF, Cao JJ, Wang B, Liang HQ, Zheng HS, Xie YL, Tap J, Lepage P, Bertalan M, Batto JM, Hansen T, Le Paslier D, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu HM, Yu C, Li ST, Jian M, Zhou Y, Li YR, Zhang XQ, Li SG, Qin N, Yang HM, Wang J, Brunak S, Dore J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Bork P, Ehrlich SD, Wang J, MetaHIT Consortium.** 2010. A human gut microbial gene catalogue established by metagenomic sequencing. Nature **464:**-U59–U70.
24. **Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA, Sogin ML.** 2007. Microbial population structures in the deep marine biosphere. Science **318:**97–100.
25. **Sogin ML, Morrison HG, Huber JA, Mark Welch DM, Huse SM, Neal PR, Arrieta JM, Herndl GJ.** 2006. Microbial diversity in the deep sea and the underexplored "rare biosphere." Proc. Natl. Acad. Sci. U. S. A. **103:**12115–12120.
26. **Gilbert JA, Steele JA, Caporaso JG, Steinbruck L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I, Somerfield P, Fuhrman JA, Field D.** 2012. Defining seasonal marine microbial community dynamics. ISME J. **6:**298–308.
27. **Edwards RA, Rodriguez-Brito B, Wegley L, Haynes M, Breitbart M, Peterson DM, Saar MO, Alexander S, Alexander EC, Rohwer F.** 2006. Using pyrosequencing to shed light on deep mine microbial ecology. BMC Genomics **7:**-57.
28. **Leininger S, Urich T, Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C.** 2006. Archaea predominate among ammonia-oxidizing prokaryotes in soils. Nature **442:**806–809.
29. **Roesch LF, Fulthorpe RR, Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW.** 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. ISME J. **1:**283–290.
30. **Bates ST, Berg-Lyons D, Caporaso JG, Walters WA, Knight R, Fierer N.** 2011. Examining the global distribution of dominant archaeal populations in soil. ISME J. **5:**908–917.
31. **Pilloni G, Granitsiotis MS, Engel M, Lueders T.** 2012. Testing the limits of 454 pyrotag sequencing: reproducibility, quantitative assessment and comparison to T-RFLP fingerprinting of aquifer microbes. PLoS One **7:**e40467. http://dx.doi.org/10.1371/journal.pone.0040467.
32. **Lin X, McKinley J, Resch CT, Kaluzny R, Lauber CL, Fredrickson J, Knight R, Konopka A.** 2012. Spatial and temporal dynamics of the microbial community in the Hanford unconfined aquifer. ISME J. **6:**1665–1676.
33. **Caporaso JG, Lauber CL, Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R.** 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. ISME J. **6:**1621–1624.
34. **Human Microbiome Project Consortium.** 2012. A framework for human microbiome research. Nature **486:**215–221.
35. **Human Microbiome Project Consortium.** 2012. Structure, function and diversity of the healthy human microbiome. Nature **486:**207–214.
36. **Frias-Lopez J, Shi Y, Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF.** 2008. Microbial community gene expression in ocean surface waters. Proc. Natl. Acad. Sci. U. S. A. **105:**3805–3810.
37. **Mackelprang R, Waldrop MP, DeAngelis KM, David MM, Chavarria KL, Blazewicz SJ, Rubin EM, Jansson JK.** 2011. Metagenomic analysis of a permafrost microbial community reveals a rapid response to thaw. Nature **480:**368–371.
38. **Mason OU, Hazen TC, Borglin S, Chain PS, Dubinsky EA, Fortney JL, Han J, Holman HY, Hultman J, Lamendella R, Mackelprang R, Malfatti S, Tom LM, Tringe SG, Woyke T, Zhou J, Rubin EM, Jansson JK.** 2012. Metagenome, metatranscriptome and single-cell sequencing reveal microbial response to deepwater Horizon oil spill. ISME J. **6:**1715–1727.
39. **Qin J, Li Y, Cai Z, Li S, Zhu J, Zhang F, Liang S, Zhang W, Guan Y, Shen D, Peng Y, Zhang D, Jie Z, Wu W, Qin Y, Xue W, Li J, Han L, Lu D, Wu P, Dai Y, Sun X, Li Z, Tang A, Zhong S, Li X, Chen W, Xu R, Wang M, Feng Q, Gong M, Yu J, Zhang Y, Zhang M, Hansen T, Sanchez G, Raes J, Falony G, Okuda S, Almeida M, LeChatelier E, Renault P, Pons N, Batto J-M, Zhang Z, Chen H, Yang R, Zheng W, Li S, Yang H, Wang J, Ehrlich SD, Nielsen R, Pedersen O, Kristiansen K, Wang J.** 2012. A metagenome-wide association study of gut microbiota in type 2 diabetes. Nature **490:**55–60.
40. **Shi Y, Tyson GW, DeLong EF.** 2009. Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. Nature **459:**266–269.
41. **Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM.** 2005. Comparative metagenomics of microbial communities. Science **308:**554–557.
42. **Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers YH, Smith HO.** 2004. Environmental genome shotgun sequencing of the Sargasso Sea. Science **304:**66–74.
43. **Weinstock GM.** 2012. Genomic approaches to studying the human microbiota. Nature **489:**250–256.
44. **Chase JM, Myers JA.** 2011. Disentangling the importance of ecological niches from stochastic processes across scales. Philos. Trans. R. Soc. Lond. B Biol. Sci. **366:**2351–2363.
45. **Chase JM.** 2007. Drought mediates the importance of stochastic community assembly. Proc. Natl. Acad. Sci. U. S. A. **104:**17430–17434.
46. **Chase JM.** 2010. Stochastic community assembly causes higher biodiversity in more productive environments. Science **328:**1388–1391.
47. **Kraft NJ, Comita LS, Chase JM, Sanders NJ, Swenson NG, Crist TO, Stegen JC, Vellend M, Boyle B, Anderson MJ, Cornell HV, Davies KF,**

Freestone AL, Inouye BD, Harrison SP, Myers JA. 2011. Disentangling the drivers of β diversity along latitudinal and elevational gradients. Science **333:**1755–1758.
48. **He Z, Xu M, Deng Y, Kang S, Kellogg L, Wu L, van Nostrand JD, Hobbie SE, Reich PB, Zhou J.** 2010. Metagenomic analysis reveals a marked divergence in the structure of belowground microbial communities at elevated CO_2. Ecol. Lett. **13:**564–575.
49. **Zhou JZ, Xue K, Xie JP, Deng Y, Wu LY, Cheng XH, Fei SF, Deng SP, He ZL, Van Nostrand JD, Luo YQ.** 2012. Microbial mediation of carbon-cycle feedbacks to climate warming. Nat. Clim. Change **2:**106–110.
50. **Zhou J, Wu L, Deng Y, Zhi X, Jiang YH, Tu Q, Xie J, Van Nostrand JD, He Z, Yang Y.** 2011. Reproducibility and quantitation of amplicon sequencing-based detection. ISME J. **5:**1303–1313.
51. **Pinto AJ, Raskin L.** 2012. PCR biases distort bacterial and archaeal community structure in pyrosequencing datasets. PLoS ONE **7:**e43093. http://dx.doi.org/10.1371/journal.pone.0043093.
52. **Zhou J, Kang S, Schadt CW, Garten CT, Jr.** 2008. Spatial scaling of functional gene diversity across various microbial taxa. Proc. Natl. Acad. Sci. U. S. A. **105:**7768–7773.
53. **Green JL, Plotkin JB.** 2007. A statistical theory for sampling species abundances. Ecol. Lett. **10:**1037–1045.
54. **Plotkin JB, Muller-Landau HC.** 2002. Sampling the species composition of a landscape. Ecology **83:**3344–3356.
55. **Chisholm RA.** 2007. Sampling species abundance distributions: resolving the veil-line debate. J. Theor. Biol. **247:**600–607.
56. **Dewdney AK.** 1998. A general theory of the sampling process with applications to the "veil line." Theor. Popul. Biol. **54:**294–302.
57. **Quince C, Curtis TP, Sloan WT.** 2008. The rational exploration of microbial diversity. ISME J. **2:**997–1006.
58. **Krebs CJ.** 1999. Ecological methodology, 2nd ed. Benjamin/Cummings, Menlo Park, CA.
59. **Alonso D, McKane AJ.** 2004. Sampling Hubbell's neutral theory of biodiversity. Ecol. Lett. **7:**901–910.
60. **Pielou EC.** 1969. An introduction to mathematical ecology. Wiley Interscience, New York, NY.
61. **Talley NJ, Fodor AA.** 2011. Bugs, stool, and the irritable bowel syndrome: too much is as bad as too little? Gastroenterology **141:**1555–1559.
62. **Lemos LN, Fulthorpe RR, Roesch LF.** 2012. Low sequencing efforts bias analyses of shared taxa in microbial communities. Folia Microbiol. (Praha) **57:**409–413.
63. **Bartram AK, Lynch MD, Stearns JC, Moreno-Hagelsieb G, Neufeld JD.** 2011. Generation of multimillion-sequence 16S rRNA gene libraries from complex microbial communities by assembling paired-end Illumina reads. Appl. Environ. Microbiol. **77:**3846–3852.
64. **Kauserud H, Kumar S, Brysting AK, Nordén J, Carlsen T.** 2012. High consistency between replicate 454 pyrosequencing analyses of ectomycorrhizal plant root samples. Mycorrhiza **22:**309–315.
65. **Mao Y, Yannarell AC, Mackie RI.** 2011. Changes in N-transforming archaea and bacteria in soil during the establishment of bioenergy crops. PLoS One **6:**e24750. http://dx.doi.org/10.1371/journal.pone.0024750.
66. **Xu LH, Ravnskov S, Larsen J, Nicolaisen M.** 2011. Influence of DNA extraction and PCR amplification on studies of soil fungal communities based on amplicon sequencing. Can. J. Microbiol. **57:**1062–1066.
67. **Schloss PD, Gevers D, Westcott SL.** 2011. Reducing the effects of PCR amplification and sequencing artifacts on 16S rRNA-based studies. PLoS One **6:**e27310. http://dx.doi.org/10.1371/journal.pone.0027310.
68. **Berry D, Ben Mahfoudh K, Wagner M, Loy A.** 2011. Barcoded primers used in multiplex amplicon pyrosequencing bias amplification. Appl. Environ. Microbiol. **77:**7846–7849.
69. **Engelbrektson A, Kunin V, Wrighton KC, Zvenigorodsky N, Chen F, Ochman H, Hugenholtz P.** 2010. Experimental factors affecting PCR-based estimates of microbial species richness and evenness. ISME J. **4:**642–647.
70. **Gomez-Alvarez V, Teal TK, Schmidt TM.** 2009. Systematic artifacts in metagenomes from complex microbial communities. ISME J. **3:**1314–1317.
71. **Huse SM, Huber JA, Morrison HG, Sogin ML, Welch DM.** 2007. Accuracy and quality of massively parallel DNA pyrosequencing. Genome Biol. **8:**R143.
72. **Kunin V, Engelbrektson A, Ochman H, Hugenholtz P.** 2010. Wrinkles in the rare biosphere: pyrosequencing errors can lead to artificial inflation of diversity estimates. Environ. Microbiol. **12:**118–123.
73. **Cam E, Nichols JD, Hines JE, Sauer JR, Jara A, Flather CH.** 2002. Disentangling sampling and ecological explanations underlying species-area relationships. Ecology **83:**1118–1130.
74. **Luo Y, Wan S, Hui D, Wallace LL.** 2001. Acclimatization of soil respiration to warming in a tall grass prairie. Nature **413:**622–625.