environmental microbiology

Environmental Microbiology (2020) **22**(6), 2140–2149 doi:10.1111/1462-2920.14973

Random sampling in metagenomic sequencing leads to overestimated spatial scaling of microbial diversity

Qichao Tu *

Institute of Marine Science and Technology, Shandong University, Qingdao, China.

Summary

Revealing the spatial scaling patterns of microbial diversity is of special interest in microbial ecology. One critical question is whether the observed spatial turnover rate truly reflect the actual spatial patterns of extremely diverse microbial communities. Using simulated mock communities, this study suggested that the currently observed microbial spatial turnover rates were overestimated by random sampling processes associated with high-throughput metagenomic sequencing. The observed *z* values were largely contributed by accumulated microbial taxa due to cumulative number of samples. This is a crucial issue because microbial communities already have very low spatial turnover rate due to the small size and potential cosmopolitism nature of microorganisms. Further investigations suggested a linear relationship between the observed and expected *z* values, which can be applied to remove random sampling noises from the observed *z* values. Adjustment of *z* values for data sets from six American forests showed much lower spatial turnover rate than that before adjustment. However, the patterns of *z* values among these six forests remained unchanged. This study suggested that our current understanding of microbial taxa–area relationships could be inaccurate. Therefore, cautions and efforts should be made for more accurate estimation and interpretation of microbial spatial patterns.

Introduction

Examining whether microbial communities follow similar ecological laws as macro-communities is challenging but of critical importance in microbial ecology (Prosser *et al*., 2007). One such example is taxa–area relationship (TAR), for which a positive power law relationship is expected between the number of species and the size of sampling area. Relating TARs with the complex microbial communities will not only provide insights into the fundamental processes that determine biodiversity (Meyer *et al*., 2018), but also help extend the generality of empirical patterns and support mechanistic hypotheses that living entities follow universal laws (Martiny *et al*., 2006; Storch *et al*., 2012). Although the relationship for macro-communities was noticed as early as in 1920s (Arrhenius, 1921), it was only about 15 years ago when recognition of TAR for microbial communities was achieved owing to the advances in molecular techniques (Green *et al*., 2004; Horner-Devine *et al*., 2004; Bell *et al*., 2005; Smith *et al*., 2005). Recent in-laboratory experiments also demonstrated evidences that microbial communities did follow this classic ecological theorem (Delgado-Baquerizo *et al*., 2018).

In contrast to TAR analyses for macro-communities, one critical issue in exploring the spatial patterns of microbial diversity is the inability to completely capture all microbial taxa in the sampling area due to the high diversity and small size of microorganisms. As a result, the observed microbial spatial turnover rate (*z* value) is often based on severely underestimated microbial taxa richness in the sampling area, especially when techniques like clone library sequencing and denaturing gradient gel electrophoresis were used. Over the past decade, high-throughput metagenomic technologies have revolutionized microbial community studies (Poisot *et al*., 2013; Zhou *et al*., 2015). Recent studies using high-throughput technologies have demonstrated the success of applying such technologies in analyzing the spatial patterns of microbial diversity by obtaining far more microbial taxa than previously (Zhou *et al*., 2008; Tripathi *et al*., 2014; Zinger *et al*., 2014; Tu *et al*., 2016; Zhou *et al*., 2016; Deng *et al*., 2018). The problem caused by rare species not detectable by traditional molecular techniques seemed to be more or less overcome. Comparison of DNA fingerprinting and meta-barcoding approaches suggests that the latter provides significantly higher and more accurate estimates of soil bacterial TAR (Terrat *et al*., 2015).

Received 27 September, 2019; revised 22 February, 2020; accepted 1 March, 2020. *For correspondence. E-mail tuqichao@outlook.com; Tel. 86 0531 88366287.

However, recent studies also suggest low reproducibility due to random sampling artefacts in metagenomic sequencing, even among technical replicates, leading to overestimated microbial β-diversity (Zhou *et al*., 2011; Zhou *et al*., 2013; Zhan *et al*., 2014). As microbial TARs can be considered as a different form of β-diversity, the spatial turnover rate of microbial communities may also be affected by random sampling processes. Specifically, each gram of soil contains as many as 10^8 prokaryotic cells and 10^4 species (Whitman *et al*., 1998; Torsvik and Øvreås, 2002; Daniel, 2005). Random sampling processes associated with metagenomic sequencing (e.g. sample collection, DNA extraction, polymerase chain reaction [PCR] amplification, library construction and sequencing) hamper complete capturing such highly diverse microbial communities. Every time a sample is sequenced at a specific depth (e.g. 30 000 reads), a number of new microbial species/operational taxonomic units (OTUs) are obtained even though these samples harbor homogeneous microbial communities in reality. As a result, noises are added to microbial spatial scaling analysis, leading to inaccurate spatial turnover rate estimation.

In this study, we aimed to investigate how microbial spatial scaling patterns were affected by random sampling issues associated with high-throughput metagenomic sequencing. We constructed simulated microbial community pools following log-normal distribution and evaluated the effects of random sampling issues on microbial TAR analyses. We hypothesized that similar to microbial β-diversity estimation, microbial spatial turnover rates are also routinely overestimated due to random sampling issues. The relationship between observed and expected *z* values were then analyzed using simulated mock communities with presetting microbial spatial scaling patterns. Effort was made to remove random sampling noises from observed microbial spatial turnover rates. The results suggested that the currently observed microbial spatial scaling patterns could be overestimated by as high as 50%. This study challenges the conventional idea that microbial spatial turnover rates are underestimated due to undetected rare species in the environment.

Results

Assessment of random sampling artefacts associated with microbial spatial scaling analyses

For macroorganisms with large body size, species–area relationship mainly refers to the relationship between the number of species found in an area and the size of the area. The number of species increases with the area with a power law relationship. In contrast, for microbial communities, it is impossible to fully capture all microorganisms in the sampling area with current technologies. Therefore, the observed microbial TAR is in fact a result of the new microbial taxa found in new sampling area (expected TAR), as well as those detected by cumulative number of samples (i.e. the Collector curve), the latter of which is basically a result of random sampling processes associated with metagenomic sequencing. Here, the Collector curve describes the number of accumulated microbial taxa with cumulative number of samples. This in general equals to sample-based species accumulative curves (Gotelli and Colwell, 2001). Due to the small size, high diversity and potential cosmopolitanism nature of microorganisms, such random sampling processes in metagenomic sequencing could have non-negligible effects on microbial TAR analyses and may even serve as the major contributor to the observed microbial TARs.

To assess the effects of random sampling processes on microbial TARs, a simulated experiment was carried out. In this simulated experiment, a typical sampling scheme with six-circled sampling regions ranging from 1 to 100 m in radius was established (Fig. 1A). Microbial communities in these sampling regions were set homogeneous so that a spatial turnover rate of zero was expected, though this can hardly be achieved in reality. However, the extremely high diversity of microbial communities hampered complete capturing of microbial taxa in the ecosystem via any high-throughput technologies, including metagenomic sequencing. Random sampling issues were associated with every procedure in surveying microbial communities, such as sampling, DNA extraction, PCR amplification, library construction and sequencing. For instance, a typical microbial community in a gram of soil contains as high as 10^4 species and 10^8 organisms (Whitman *et al*., 1998; Torsvik and Øvreås, 2002; Daniel, 2005) and usually follows log-normal species abundance distribution (Shoemaker *et al*., 2017; Wu *et al*., 2019). When 30 000 high-quality sequences per sample were obtained, simulated data sets suggested that the number of captured microbial taxa per sample was about 4500. The more samples that were sequenced, the more microbial taxa were captured. Therefore, the observed microbial TAR (Fig. 1B and C) could be in fact an analogue of "the Collector curve" (i.e. sample-based taxa accumulation curve) (Fig. 1D and E). That being said, the observed microbial TARs were more likely describing the relationship between the number of accumulated microbial taxa with cumulative number of samples rather than sampling area. It is therefore of critical importance to figure out to what level microbial spatial scaling patterns are affected by random sampling processes. And what is more important is how to minimize such random sampling effects on microbial spatial turnover rates.

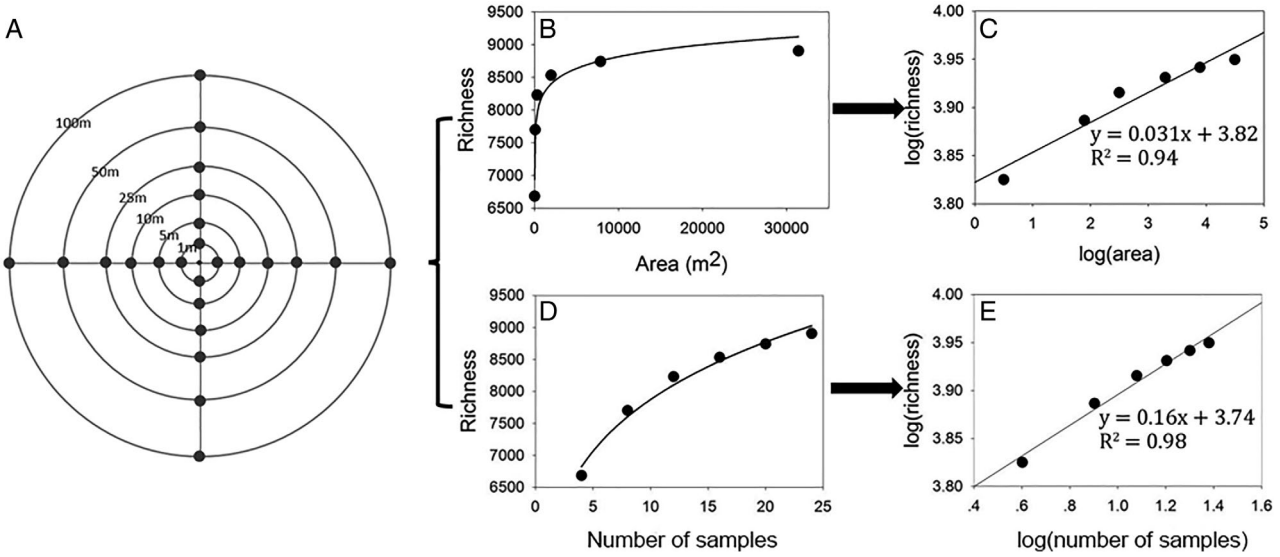

Fig. 1. Comparative illustration of microbial TAR and the Collector curve.
A. A typical microbial TAR sampling design.
B. Relationship between species richness and sampling area.
C. Relationship between log transformed species richness and sampling area.
D. Relationship between species richness and the number of samples.
E. Relationship between log transformed species richness and the number of samples.

Random sampling processes led to dramatically overestimated microbial spatial turnover rates

To evaluate to what degree random sampling processes affected microbial spatial turnover rates, a simulated microbial TAR experiment was designed and performed that each new sampling area had 500 new taxa comparing to the inner sampling region. Mock community pools in each new sampling area were composed of 10^4 taxa and 10^8 organisms in each gram of soil. For each new sampling area, four samples were collected. For each sample, a sequencing depth of 30 000 high-quality reads was mimicked via random sampling from the corresponding community pools. Under such circumstances, a theoretical microbial spatial turnover rate of 0.025 was expected (Fig. 2A). However, the observed microbial spatial turnover rate turned out to be 0.052 (Fig. 2B). This was more than doubled comparing to the theoretical value. Notably, it was also found that the expected *z* value (0.025, Fig. 2A) cannot be obtained by simply subtracting the "noised" slope value (0.031, Fig. 1E) from the observed *z* value (0.052, Fig. 2B).

Evaluation of experimental approaches to reduce random sampling issues in microbial spatial scaling analyses

Random sampling issues occur when insufficient sampling effort is made towards the studying object. Therefore, two approaches are expected to reduce random sampling noises in microbial spatial scaling analyses, including increasing the sequencing depth and increasing the number of samples. Simulated microbial TAR experiments with homogeneous communicates were set up to evaluate the effects of sequencing depth and number of samples in reducing random sampling issues. Under such circumstances, *z* value of zero was expected for microbial TAR. Therefore, the observed microbial TAR slope coefficient values solely represented noises introduced by random sampling processes (z_{rs}). We first evaluated how increasing number of samples reduced random sampling noises in microbial spatial scaling analyses. A typical sequencing depth of 30 000 high-quality sequences was generated for each sample. The relationship between z_{rs} and number of samples was investigated. As expected, the z_{rs} value decreased with increasing number of samples (Fig. 3A). When 10 samples were collected for each sampling area, a z_{rs} value of 0.0144 could be observed. Secondly, the relationship between z_{rs} and sequencing depth was investigated. For each sampling area, a typical sampling size of four samples was made. Again, the value of z_{rs} decreased with increasing sequencing depth (Fig. 3B). When a sequencing depth of 100 000 was reached, the value of z_{rs} decreased to 0.0154. Combined approach with 10 samples per sampling area and sequencing depth of 100 000 sequences per sample was also evaluated. As a result, the value of z_{rs} decreased to 0.0085. Although much lower, this was still a high noise in microbial TAR

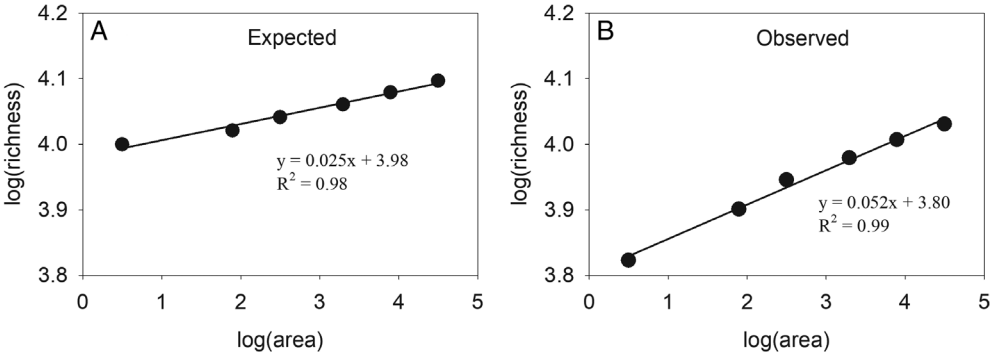

Fig. 2. Expected (A) and observed microbial TARs (B) when (i) each new sampling area was associated with 500 new species; (ii) sequencing depth of 30 000 sequences/sample was mimicked for each mock community; and (iii) each mock community pool was composed of 10^4 species and 10^8 organisms.

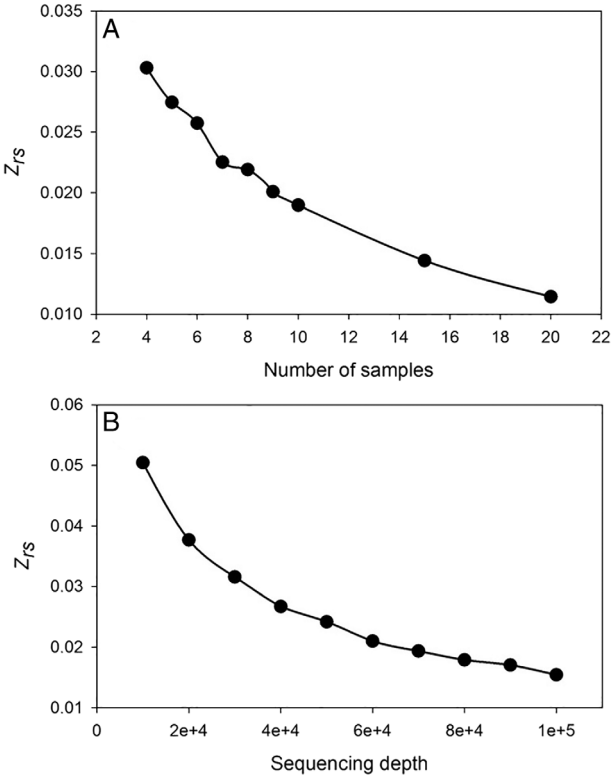

Fig. 3. Relationship between z_{rs} (slope of microbial TARs contributed by random sampling processes) and number of samples (A) and sequencing depth (B).
As expected, decreased z_{rs} values with increasing number of samples and sequencing depth could be observed.

analyses because microbial communities usually had flat slopes.

Computational adjustment of overestimated microbial spatial turnover rates

Effort was then made to more effectively reduce random sampling noises from microbial spatial scaling analyses mathematically and computationally. Theoretically, the observed microbial spatial turnover rate sources from two parts including the actual *z* value and the noises introduced by random sampling processes (z_{rs}). However, due to current technical limitations that the complex microbial communities cannot be fully captured, the actual spatial turnover rate is only partially reflected in the observed *z* value. Therefore, the relationship between the observed *z* value (z_{obs}), the expected z value (z_{exp}) and the random sampling noises (z_{rs}) can be described as the following function:

$$z_{obs} = kz_{exp} + z_{rs}, \quad (1)$$

where z_{obs} and z_{exp}, respectively, represent the observed and expected microbial spatial turnover rates, z_{rs} is the noises introduced by random sampling processes and *k* is the factor referring to the degree that the actual microbial spatial turnover rate that can be captured under current sequencing depth.

Simulated microbial TAR experiments were designed to verify the relationship between the observed and expected *z* values, with the purpose to gain clues for correct estimation of the actual microbial spatial turnover rates in the ecosystem. Because sequencing depth is a major factor related with random sampling processes, experiments analyzing the relationship between microbial spatial scaling patterns and sequencing depth were then designed. Five sequencing depths varying from 10 000 to 50 000 sequences per sample were simulated. For each increased sampling area, an addition of 500 to 3000 new microbial taxa was set. The relationship between the observed and expected *z* values was then inspected (Equation [1]). As a result, clear linear relationship could be observed between the observed and expected *z* values (Fig. 4A–E). Such relationship was consistent with the linear function inferred theoretically for the observed microbial spatial scaling patterns. This

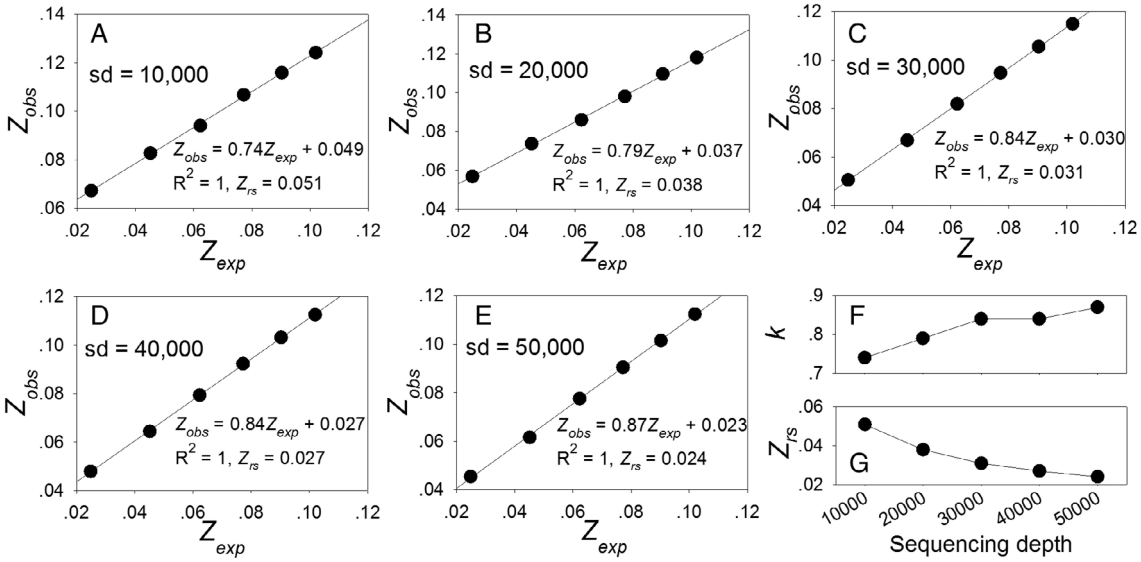

Fig. 4. A–E. The relationship between observed and expected z values under different sequencing depth.
F,G. *k* and z_{rs} values under different sequencing depth.
For each sequencing depth, microbial TARs were preset at six degrees, including 500, 1000, 1500, 2000, 2500 and 3000 new species per new sampling area; sd denotes sequencing depth.

demonstrated that the observed microbial spatial turnover rate was determined by two factors, including the "noised" *z* value due to random sampling and the "captured" *z* value under current sequencing depth. Notably, increased *k* (Fig. 4F) and decreased z_{rs} (Fig. 4G) values with increasing sequencing depth could be respectively observed. Such trends of *k* and z_{rs} values were within expectation because increased sequencing depth could result in more captured microbial taxa in the environment, thus reducing random sampling effects in microbial TAR analyses.

Adjusting z values for real studies

The above *z* adjustment equation was then applied to adjust the observed *z* values in six forests in North and Central America. For each forest, four samples were, respectively, collected at radii of 1, 10, 50, 100 and 200 m, resulting in 20 samples per forest. For each sample, a random subsampling of 20 000 sequences per sample was performed. Microbial TARs were analyzed at an identity cut-off of 97% for OTU clustering. Estimated *k* and z_{rs} values under this sampling design and sequencing depth were, respectively, 0.77 and 0.031. The observed *z* values without adjustment were 0.068 for Barro Colorado Island, Panama (BCI), 0.050 for Luquillo Long-Term Ecological Research (LTER), Puerto Rico (LUQ), 0.056 for Coweeta LTER, North Carolina (CWT), 0.053 for H.J. Andrews LTER, Oregon (AND), 0.054 for Harvard Forest LTER, Massachusetts (HFR) and 0.053 for Niwot Ridge LTER, Colorado (NWT) respectively.

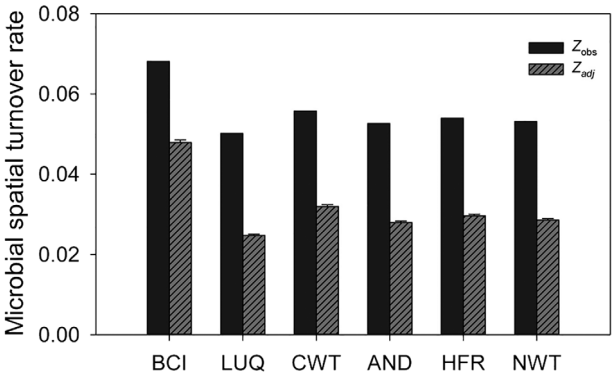

Fig. 5. Observed and adjusted microbial TAR *z* values in six American forests based on 16S amplicon sequencing data sets.
Error bars represent standard deviations.
AND, H.J. Andrews LTER, Oregon; BCI, Barro Colorado Island, Panama; CWT, Coweeta LTER, North Carolina; HFR, Harvard Forest LTER, Massachusetts; LUQ, Luquillo LTER, Puerto Rico; NWT: Niwot Ridge LTER, Colorado.

After adjustment, the adjusted *z* values were 0.048 for BCI, 0.025 for LUQ, 0.032 for CWT, 0.028 for AND, 0.030 for HFR and 0.029 for NWT respectively. These adjusted values were approximately 50% to 70% of the observed values (Fig. 5). Notably, it was noticed that the patterns of variations of observed and adjusted *z* values among different forests remained unchanged. This suggested that *z* adjustment resulted in more accurate microbial spatial turnover rates but did not change the patterns of *z* value variations among different sampling sites.

Discussion

Random sampling issues are critical for microbial spatial scaling analyses

Determining whether microbial communities follow general ecological laws (e.g. TAR) is of critical importance in microbial ecology. This not only provides insights into the fundamental processes that determine biodiversity (Meyer *et al*., 2018), but also helps extend the generality of empirical patterns and support mechanistic hypotheses that living entities follow universal laws (Martiny *et al*., 2006; Storch *et al*., 2012). As one of the potential ecological laws also followed by microorganisms, spatial patterns of microbial diversity have gained extensive foci in microbial ecology. However, owing to the huge diversity of microbial communities in natural ecosystems (Whitman *et al*., 1998; Torsvik and Øvreås, 2002; Daniel, 2005), it is almost impossible to thoroughly investigate microbial spatial scaling patterns as what have been done for plants and animals (Connor and McCoy, 1979).

Random sampling is an unneglectable issue when surveying the immense microbial world, no matter which technology is used. In the old time using traditional molecular techniques, microbial spatial turnover rates might be underestimated because the effects of rare taxa could be much greater than that of random sampling processes, especially when only a few dozens or hundreds of microbial taxa were captured. However, in the high-throughput metagenomic sequencing era when rare microbial taxa become detectable, random sampling could be a major issue affecting microbial diversity analysis, such as overestimating microbial β-diversity (Zhou *et al*., 2011; Zhou *et al*., 2013; Zhan *et al*., 2014). This is an especially critical issue for microbial communities having relatively flat spatial scaling slopes if random sampling noises exceed the actual microbial spatial turnover rates. It is therefore of necessity to investigate the effects of random sampling noises on microbial spatial scaling analyses and to remove such noises from the observed microbial spatial turnover rates.

Methods to reduce random sampling noises and our recommendations

Interestingly, this study found that the effects of random sampling noises on microbial spatial scaling patterns were mainly determined by sequencing depth, which was also the factor affecting the degree to which the actual microbial spatial turnover rate could be captured. A linear relationship was observed between the observed and expected microbial spatial turnover rates. A mathematical approach was developed to adjust microbial *z* values by removing random sampling noises. Based on the linear function, random sampling affected microbial *z* values in two parts, including z_{rs} (noises introduced by random sampling) and *k* (the degree of actual *z* values captured by current sequencing depth). Importantly, both z_{rs} and *k* were mainly determined by sequencing depth. That being said, in a community with known microbial diversity (e.g. the number of microbial species and organisms in each gram of soil) and fixed experimental design (e.g. number and size of sampling areas and the number of samples per sampling area), removing random sampling noises from microbial spatial scaling patterns could be approached mathematically (i.e. by estimating z_{rs} and *k* values). Re-estimation of microbial spatial turnover rates suggested that *z* values for soil microbial communities ranged from 0.025 to 0.048 in six American forests. Importantly, the patterns of adjusted *z* values among different forests remained unchanged, suggesting that comparative results of microbial spatial scaling patterns among different ecosystems were still valid.

Random sampling noises could also be reduced experimentally, such as by increasing the number of samples and sequencing depth, which are also evaluated in this study. Previous studies also suggested that the slope coefficients of microbial TARs were influenced by several different factors including removing rare taxa or not, the number of sampling sites (Zinger *et al*., 2014) and the cut-off used for OTU clustering (Horner-Devine *et al*., 2004). Among these factors, increasing the number of samples could reduce random sampling effects without artificially changing microbial community structure and/or microbial taxa definition. Besides that, increasing sequencing depth and sampling area are also expected to experimentally reduce random sampling noises. However, these approaches seem neither economical nor practical. For instance, using homogeneous mock communities, z_{rs} values of 0.024 and 0.019 were still, respectively, observed when increasing sequencing depth to 50 000 and number of samples to 10. Based on the findings in this study, we therefore recommend to perform microbial spatial scaling studies with reasonable sequencing depth (e.g. ≥50 000 sequences per sample) and number of samples (e.g. ≥4 samples per increasing sampling area), then to remove random sampling noises computationally as described in this study.

In addition to using nested sampling design (Green *et al*., 2004; Horner-Devine *et al*., 2004; Zhou *et al*., 2008; Tu *et al*., 2016; Deng *et al*., 2018), microbial spatial patterns can also be analyzed using non-nested sampling designs (Bell *et al*., 2005; Zinger *et al*., 2014). Theoretically, there is no difference between these two sampling designs in investigating microbial spatial scaling patterns. However, it should be noted that the number of samples per increasing sampling area are usually the same for nested sampling designs, while it may not be the case for non-nested sampling designs. This could be

a potential issue in estimating and removing random sampling noises from microbial spatial scaling patterns because uneven number of samples could lead to uneven random sampling noises. Similarly, the sequencing depth, another factor related with random sampling noises, should also be the same among different samples.

The actual situation of microbial TAR may still remain unknown

We also would like to point out that the proposed *z* adjustment method only aimed to remove noises caused by random sampling issues associated with metagenomic sequencing, but not to uncover the genuine microbial TARs in the ecosystems. In fact, uncovering the genuine microbial TARs is almost impossible based on our current knowledge and technologies. Assuming a typical microbial TAR experiment was carried out in a small region (e.g. 100 m in radius) and the top 20 cm soil was collected, and assuming each gram of soil contained 10^8 cells, the total microbial cells in this small region should be $\sim 6.28 \times 10^{17}$. This makes both experimental surveying and computational simulation of such highly diverse microbial communities almost impossible. Therefore, the genuine situation of microbial TARs in natural ecosystems still remains as an enigma.

Although many studies have been carried out showing different microbial TARs for different taxonomic groups in different ecosystems (Green *et al*., 2004; Horner-Devine *et al*., 2004; Bell *et al*., 2005; Green and Bohannan, 2006; Zhou *et al*., 2008; Zinger *et al*., 2014; Tu *et al*., 2016; Deng *et al*., 2018), there are also arguments questioning the existence of microbial TARs (Finlay, 2002; Fenchel and Finlay, 2005; Green and Bohannan, 2006). Due to their small size and high abundance, microorganisms may disperse further and faster, resulting in almost nonexistent dispersal limitation and cosmopolitan distributions. Intriguingly, recent studies by deep sequencing suggested a persistent microbial seed bank in the global ocean and the Western English Channel (Caporaso *et al*., 2012; Gibbons *et al*., 2013), supporting the arguments that microbial TARs may not exist. That being said, the currently observed microbial TARs could be solely due to random sampling issues. In this study, we are not joining the debate whether microbial TARs exist or not. Rather, we show evidences that the currently observed microbial spatial turnover rates via high-throughput metagenomic sequencing technologies are overestimated by random sampling processes. This partially complies with these recent discoveries (Caporaso *et al*., 2012; Gibbons *et al*., 2013), but seems to be inconsistent with Woodcock et al.'s (2006) viewpoint that undetected rare species underestimated microbial TARs.

Further implications

In addition to the spatial scaling patterns that we analyzed in this study, random sampling issues associated with high-throughput metagenomic sequencing are also expected to affect other microbial diversity studies, such as taxa–time relationship, beta diversity, distancedecay relationships (DDRs) and taxa–abundance distributions (TADs). Among these, the effects of random sampling issues on taxa–time relationship for microbial communities are similar to what have been investigated for microbial spatial scaling patterns in this study. Beta diversity of microbial communities are also expected to be overestimated by random sampling processes, as previously reported (Zhou *et al*., 2011; 2013; Zhan *et al*., 2014). Although random sampling issues may lead to overestimated beta diversity, the effects of random sampling on DDRs and TADs of microbial communities are still not clear. For instance, when pairwise community similarities were calculated for microbial DDRs, the slopes of microbial DDRs may not be strongly affected by random samplings issues. Therefore, further investigations are needed to figure out how random sampling issues affect microbial DDRs and TADs.

In conclusion, determining the patterns of microbial diversity across space and time is a central issue in microbial ecology. However, revealing such patterns is difficult due to the immense diversity of microbial communities. This study showed evidences that the currently observed microbial spatial turnover rates via high-throughput metagenomic sequencing were overestimated due to random sampling processes. This is such a critical issue that conventional experimental procedures such as standardized sampling and increased number of samples and sequencing depth cannot solve. As random sampling issues are common in microbial ecology studies, the methodological framework presented in this study may also provide valuable clues to other microbial diversity studies, such as microbial DDRs and TADs.

Methods

Methodological framework

This study aimed to investigate how random sampling issues affected microbial spatial scaling analyses and seek potential solutions to remove random sampling noises from microbial spatial turnover rates. Using mock communities and virtual microbial TAR sampling design, the relationship between random sampling noises, observed microbial *z* values and expected microbial *z* values was investigated (Fig. 6). We also evaluated different experimental methods (e.g. increasing number of samples and sequencing depth) to see to what degree random sampling noises could be removed

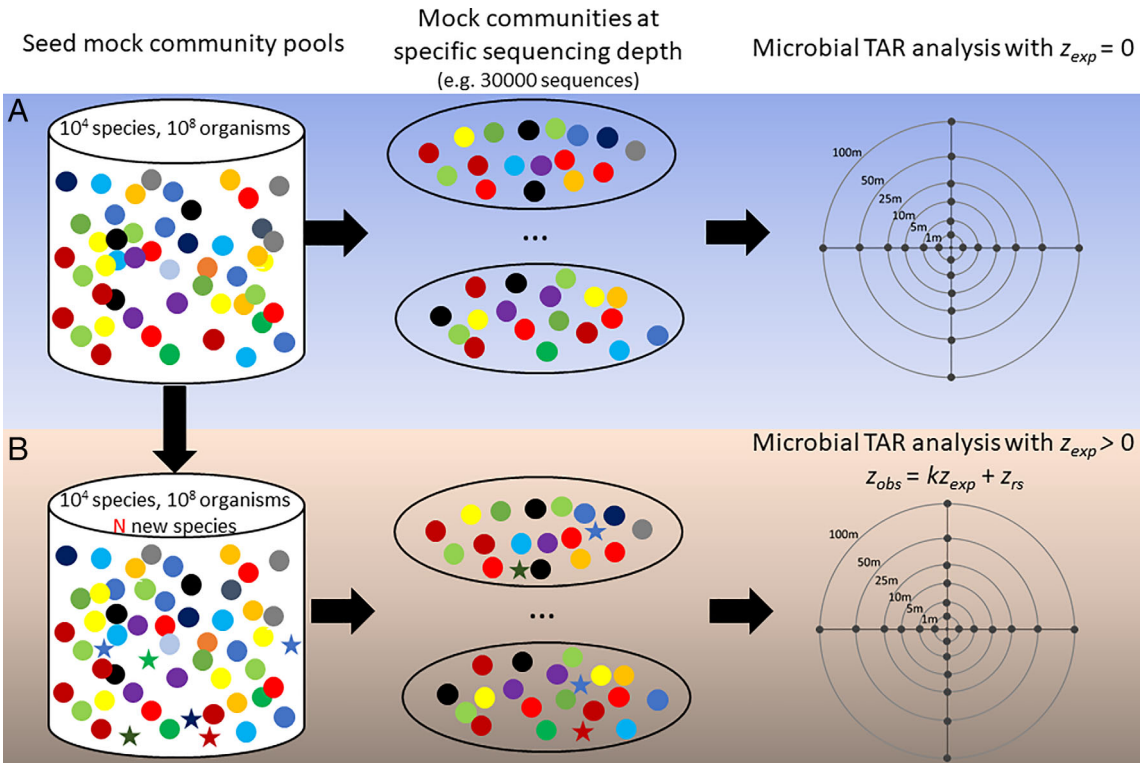

Fig. 6. Schematic illustration of simulated microbial TAR analyses in this study.
A. Microbial TAR analysis with $z_{exp} = 0$ (homogeneous community). A seed mock community pool following lognormal species abundance distribution was first constructed with 10^4 species and 10^8 organisms.
B. Microbial TAR analysis with $z_{exp} > 0$.
Mock community pools with *N* new species were constructed from the seed mock community pool. For both microbial TAR analyses, mock communities at specific sequencing depth (e.g. 30000 sequences) were constructed from the seed mock community.
Microbial TAR analyses were carried out with the virtual sampling design. The relationship between z_{obs}, z_{exp} and z_{rs} were analyzed. Different colours represent different microbial species. Star symbols represent new species. [Color figure can be viewed at wileyonlinelibrary.com]

experimentally. Effort was also made to develop computational approaches to remove random sampling noises. The computational approach was then applied to a previous microbial TAR study in six American forests. The patterns of microbial TARs among different forests were analyzed.

Seed community construction

Because metagenomic amplicon sequencing of soil microbial communities usually starts from 1 g of soil, a seed mock community pool composed of 10^4 species and 10^8 organisms (i.e. sequences) was generated (Fig. 6A). The seed mock community pool followed lognormal species abundance distribution, a pattern followed by most microbial communities in both natural and artificial ecosystems (Shoemaker *et al*., 2017; Wu *et al*., 2019). This seed mock community was later used as a reference to generate other mock community pools (Fig. 6B). The R package mobsim (Felix *et al*., 2018) was used to generate the seed mock community and other mock community pools in this study.

Evaluating random sampling effects on microbial spatial turnover rates

To evaluate how random sampling issues affected microbial spatial turnover rates, a typical virtual microbial TAR sampling scheme was designed (Fig. 6A). In this sampling scheme, multiple sampling areas were set with different radii. For each increased sampling area, four samples were collected. For each sample, a random subsampling of 30 000 sequences from the seed mock community was performed, simulating a typical metagenomic sequencing process. The R command “sample.int” was used to randomly select a subset of organisms from the seed mock community pool. An OTU abundance table was then generated for all samples in the microbial TAR experimental design. Microbial spatial scaling patterns were then analyzed using the generalization of Arrhenius’ (1921) equation, by a double logarithmic transformation:

$$\log(S_{obs}) = c + z \times \log(A),$$

where S_{obs} is the number of observed species, *c* is the intercept parameter, *A* is the area and *z* is the slope coefficient of TAR. Inspired by microbial TARs, the relationship between S_{obs} and sample numbers was also analyzed similarly. As a homogeneous composition of microbial communities was expected in this experimental design, the observed microbial spatial turnover rates should be solely a result of random sampling processes.

Testing different methods to reduce random sampling effects

The number of samples and sequencing depth were two major factors related with random sampling effects, and therefore were evaluated here. The evaluation was carried out under the assumption that microbial communities in the same sampling area were homogeneous, thus microbial *z* values of zero were expected (Fig. 6A), though this can be hardly achieved in reality. For each sampling area, 4–20 mock communities at subsampling depth of 30 000 per new sampling area were generated and subjected to microbial spatial scaling analyses. For sequencing depth, random subsampling of 10^4 to 10^5 sequences per sample was performed, while the number of samples per increased sampling area remained at four. Random subsampling effects (z_{rs} values) under different conditions were analyzed. The relationship between z_{rs} values, the number of samples and sequencing depth was investigated.

Computational adjustment of microbial z *values*

In order to figure out whether random sampling noises could be removed from microbial spatial scaling patterns computationally, mock communities with a series of presetting microbial spatial turnover rates were constructed to analyze the relationship between observed and expected microbial *z* values (Fig. 6B). The seed mock community constructed in the first step was used to represent microbial communities in the central sampling area. For each increased sampling area, mock community pools were generated based on the seed mock community. These community pools in each new sampling area were designed to have *n* new species than their inner sampling area so that a presetting microbial spatial turnover rate was followed. Six degrees of microbial spatial scaling patterns were designed that the numbers of new species in each larger sampling area were respectively 500, 1000, 1500, 2000, 2500 and 3000. To analyze the relationship between random sampling effects and sequencing depth, random subsampling of 10 000–50 000 organisms from the mock community pools were performed. OTU abundance tables were generated for all 20 samples in the TAR experimental design. The relationship between observed and expected microbial *z* values was then analyzed.

Case study

The above-developed *z* adjustment method was applied to analyze microbial TARs in six American forests. The 16S amplicon sequencing data set was downloaded according to the previous study (Zhou *et al*., 2016). These six forests included AND, CWT, HFR, LUQ, NWT and BCI. A total of 126 soil samples (0–10 cm, 21 samples per site) were collected from the six forest sites in the summer of 2012 for microbial community analysis. Soil DNA was extracted, subjected to PCR amplification and sequenced by Illumina MiSeq Platform. The sample located in the central point was discarded from analysis due to its zero area size. Microbial spatial scaling patterns were analyzed for these six soil microbial communities. A comparison of the observed and adjusted microbial *z* values was carried out. Details for experimental designing, sampling and sequence processing could be found in Zhou *et al*. (2016).

Acknowledgements

This study was supported by the National Natural Science Foundation of China (31700427, 31971446), the Zhejiang Provincial Natural Science Foundation of China (LQ17D060002), the Qilu Young Scholarship of Shandong University, the Bureau of Science and Technology of Zhoushan (2017C82218), the Open Funding of State Key Laboratory of Applied Microbiology Southern China (SKYAM002-2016) and the Open Project of Key Laboratory of Environmental Biotechnology, CAS (Grant No kf2016002). The funders had no role in study design, data collection and interpretation or the decision to submit the work for publication.

Data availability

Source codes for mock community construction, z_{rs} and *k* estimations are publicly available at https://github.com/qichao1984/TAR_RS.

References

Arrhenius, O. (1921) Species and area. *J Ecol* **9**: 95–99.

Bell, T., Ager, D., Song, J.-I., Newman, J.A., Thompson, I.P., Lilley, A.K., and Van der Gast, C.J. (2005) Larger islands house more bacterial taxa. *Science* **308**: 1884–1884.

Caporaso, J.G., Paszkiewicz, K., Field, D., Knight, R., and Gilbert, J.A. (2012) The Western English Channel contains a persistent microbial seed bank. *ISME J* **6**: 1089–1093.

Connor, E.F., and McCoy, E.D. (1979) The statistics and biology of the species-area relationship. *Am Nat* **113**: 791–833.

Daniel, R. (2005) The metagenomics of soil. *Nat Rev Microbiol* **3**: 470.

Delgado-Baquerizo, M., Eldridge, D.J., Hamonts, K., Reich, P.B., and Singh, B.K. (2018) Experimentally testing the species-habitat size relationship on soil bacteria: A proof of concept. *Soil Biol Biochem* **123**: 200–206.

Deng, Y., Ning, D., Qin, Y., Xue, K., Wu, L., He, Z., *et al*. (2018) Spatial scaling of forest soil microbial communities across a temperature gradient. *Environ Microbiol* **20**: 3504–3513.

Felix, M., Katharina, G., DJ McGlinn, X Xiao, JM Chase (2018) mobsim: An r package for the simulation and measurement of biodiversity across spatial scales. *Methods Ecol Evol* **9**: 1401–1408.

Fenchel, T., and Finlay, B.J. (2005) Bacteria and island biogeography. *Science* **309**: 1997–1999.

Finlay, B.J. (2002) Global dispersal of free-living microbial eukaryote species. *Science* **296**: 1061–1063.

Gibbons, S.M., Caporaso, J.G., Pirrung, M., Field, D., Knight, R., and Gilbert, J.A. (2013) Evidence for a persistent microbial seed bank throughout the global ocean. *Proc Natl Acad Sci USA* **110**: 4651–4655.

Gotelli, N.J., and Colwell, R.K. (2001) Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecol Lett* **4**: 379–391.

Green, J., and Bohannan, B.J. (2006) Spatial scaling of microbial biodiversity. *Trends Ecol Evol* **21**: 501–507.

Green, J.L., Holmes, A.J., Westoby, M., Oliver, I., Briscoe, D., Dangerfield, M., *et al*. (2004) Spatial scaling of microbial eukaryote diversity. *Nature* **432**: 747.

Horner-Devine, M.C., Lage, M., Hughes, J.B., and Bohannan, B.J.M. (2004) A taxa–area relationship for bacteria. *Nature* **432**: 750.

Martiny, J.B.H., Bohannan, B.J.M., Brown, J.H., Colwell, R. K., Fuhrman, J.A., Green, J.L., *et al*. (2006) Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**: 102.

Meyer, K.M., Memiaghe, H., Korte, L., Kenfack, D., Alonso, A., and Bohannan, B.J.M. (2018) Why do microbes exhibit weak biogeographic patterns? *ISME J* **12**: 1404–1413.

Poisot, T., Pequin, B., and Gravel, D. (2013) High-throughput sequencing: a roadmap toward community ecology. *Ecol Evol* **3**: 1125–1139.

Prosser, J.I., Bohannan, B.J.M., Curtis, T.P., Ellis, R.J., Firestone, M.K., Freckleton, R.P., *et al*. (2007) The role of ecological theory in microbial ecology. *Nat Rev Microbiol* **5**: 384.

Shoemaker, W.R., Locey, K.J., and Lennon, J.T. (2017) A macroecological theory of microbial biodiversity. *Nat Ecol Evol* **1**: 0107.

Smith, V.H., Foster, B.L., Grover, J.P., Holt, R.D., Leibold, M.A., and Denoyelles, F., Jr. (2005) Phytoplankton species richness scales consistently from laboratory microcosms to the world's oceans. *Proc Natl Acad Sci USA* **102**: 4393–4396.

Storch, D., Keil, P., and Jetz, W. (2012) Universal species–area and endemics–area relationships at continental scales. *Nature* **488**: 78–81.

Terrat, S., Dequiedt, S., Horrigue, W., Lelievre, M., Cruaud, C., Saby, N., *et al*. (2015) Improving soil bacterial taxa–area relationships assessment using DNA metabarcoding. *Heredity* **114**: 468.

Torsvik, V., and Øvreås, L. (2002) Microbial diversity and function in soil: from genes to ecosystems. *Curr Opin Microbiol* **5**: 240–245.

Tripathi, B.M., Lee-Cruz, L., Kim, M., Singh, D., Go, R., Shukor, N.A., *et al*. (2014) Spatial scaling effects on soil bacterial communities in Malaysian tropical forests. *Microb Ecol* **68**: 247–258.

Tu, Q., Deng, Y., Yan, Q., Shen, L., Lin, L., He, Z., *et al*. (2016) Biogeographic patterns of soil diazotrophic communities across six forests in the North America. *Mol Ecol* **25**: 2937–2948.

Whitman, W.B., Coleman, D.C., and Wiebe, W.J. (1998) Prokaryotes: the unseen majority. *Proc Natl Acad Sci* **95**: 6578–6583.

Woodcock, S., Curtis, T.P., Head, I.M., Lunn, M., and Sloan, W.T. (2006) Taxa-area relationships for microbes: the unsampled and the unseen. *Ecol Lett* **9**: 805–812.

Wu, L., Ning, D., Zhang, B., Li, Y., Zhang, P., Shan, X., *et al*. (2019) Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nat Microbiol* **4**: 1183–1195.

Zhan, A., Xiong, W., He, S., and MacIsaac, H.J. (2014) Influence of artifact removal on rare species recovery in natural complex communities using high-throughput sequencing. *PLoS One* **9**: e96928.

Zhou, J., Deng, Y., Shen, L., Wen, C., Yan, Q., Ning, D., *et al*. (2016) Temperature mediates continental-scale diversity of microbes in forest soils. *Nat Commun* **7**: 12083.

Zhou, J., He, Z., Yang, Y., Deng, Y., Tringe, S.G., and Alvarez-Cohen, L. (2015) High-throughput metagenomic technologies for complex microbial community analysis: open and closed formats. *MBio* **6**: e02288–e02214.

Zhou, J., Jiang, Y.-H., Deng, Y., Shi, Z., Zhou, B.Y., Xue, K., *et al*. (2013) Random sampling process leads to overestimation of β-diversity of microbial communities. *MBio* **4**: e00324–e00313.

Zhou, J., Kang, S., Schadt, C.W., and Garten, C.T. (2008) Spatial scaling of functional gene diversity across various microbial taxa. *Proc Natl Acad Sci* **105**: 7768–7773.

Zhou, J., Wu, L., Deng, Y., Zhi, X., Jiang, Y.-H., Tu, Q., *et al*. (2011) Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* **5**: 1303.

Zinger, L., Boetius, A., and Ramette, A. (2014) Bacterial taxa–area and distance–decay relationships in marine environments. *Mol Ecol* **23**: 954–964.