

RESEARCH ARTICLE

Assessment of microbial α -diversity in one meter squared topsoil

Shuzhen Li^{1,2}, Xiongfeng Du^{1,3}, Kai Feng^{1,3}, Yueni Wu^{1,3}, Qing He^{1,3}, Zhujun Wang^{1,3}, Yangying Liu^{1,3}, Danrui Wang^{1,3}, Xi Peng^{1,3}, Zhaojing Zhang⁴, Arthur Escalas⁵, Yuanyuan Qu², Ye Deng^{1,3,4,*}

¹ CAS Key Laboratory of Environmental Biotechnology, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing 100085, China

² Key Laboratory of Industrial Ecology and Environmental Engineering (Ministry of Education), School of Environmental Science and Technology, Dalian University of Technology, Dalian, 116024, China

³ College of Resources and Environment, University of Chinese Academy of Sciences, Beijing 100049, China

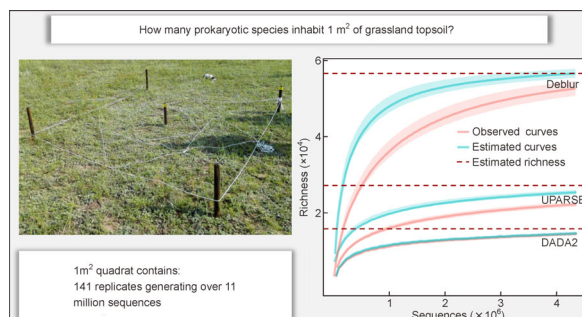
⁴ Institute for Marine Science and Technology, Shandong University, Qingdao 266237, China

⁵ MARBEC, Université de Montpellier, CNRS, IRD, IFREMER, Montpellier Cedex 5, 34090, France

HIGHLIGHTS

- Roughly 15 919 to 56 985 prokaryotic species inhabited in 1 m² grassland topsoil.
- Three clustering tools, including DADA2, UPARSE and Deblur showed huge differences.
- Nearly 500 000 sequences were required to catch 50% species.
- Insufficient sequencing depth greatly affected observed and estimated richness.
- Higher order of Hill numbers reached saturation with fewer than 100 000 sequences.

GRAPHICAL ABSTRACT



ABSTRACT

ARTICLE INFO

Article history:

Received March 7, 2021

Revised June 15, 2021

Accepted June 20, 2021

Keywords:

Grassland

Topsoil

Prokaryote

Richness

α -diversity

Hill number

Due to the tremendous diversity of microbial organisms in topsoil, the estimation of saturated richness in a belowground ecosystem is still challenging. Here, we intensively surveyed the 16S rRNA gene in four 1 m² sampling quadrats in a typical grassland, with 141 biological or technical replicates generating over 11 million sequences per quadrat. Through these massive data sets and using both non-asymptotic extrapolation and non-parametric asymptotic approaches, results revealed that roughly 15 919±193, 27 193±1076 and 56 985±2347 prokaryotic species inhabited in 1 m² topsoil, classifying by DADA2, UPARSE (97% cutoff) and Deblur, respectively, and suggested a huge difference among these clustering tools. Nearly 500 000 sequences were required to catch 50% species in 1 m², while any estimator based on 500 000 sequences would still lose about a third of total richness. Insufficient sequencing depth will greatly underestimate both observed and estimated richness. At least ~911 000, ~3 461 000, and ~1 878 000 sequences were needed for DADA2, UPARSE, and Deblur, respectively, to catch 80% species in 1 m² topsoil, and the numbers of sequences would be nearly twice to three times on this basis to cover 90% richness. In contrast, α -diversity indexes characterized by higher order of Hill numbers, including Shannon entropy and inverse Simpson index, reached saturation with fewer than 100 000 sequences, suggesting sequencing depth could be varied greatly when focusing on exploring different α -diversity characteristics of a microbial community. Our findings were fundamental for microbial studies that provided benchmarks for the extending surveys in large scales of terrestrial ecosystems.

© Higher Education Press 2021

* Corresponding author

E-mail address: yedeng@rcees.ac.cn (Y. Deng)

1 Introduction

Species richness is a fundamental measurement for community α -diversity, as well as forming the basis of many ecological models of community structure (Gotelli and Colwell, 2001). However, for soil prokaryotic assemblages (i.e., bacterial and archaeal community), it is believed that saturated richness, even approximate saturation, has never been reached due to insufficient sampling of the tremendous number of individuals (Gotelli and Colwell, 2001; Schloss and Handelsman, 2006; Roesch et al., 2007). High-throughput techniques, as sequenced 16S rRNA genes in environmental DNA, have become powerful tools to characterize this astonishing diversity of microorganisms (Roesch et al., 2007). The simplest way to measure microbial richness is to count how many species, as represented by Operational Taxonomic Units (OTUs) or Amplicon Sequence Variants (ASVs), are observed within a sample. Many studies have attempted to uncover this extraordinary diversity, especially at large, even global scales (Tedersoo et al., 2014; Delgado-Baquerizo et al., 2018; Wu et al., 2019). In contrast, little research has focused on species richness at small scales. In fact, it is more advantageous to conduct thorough sampling at small scales to compensate for the potential under-sampling bias that could result in an underestimation of richness. It can also provide a baseline value for microbial diversity in a given small area, and a theoretical basis for large-scale extrapolation.

In general, for diverse prokaryotic taxa, as more individuals are sequenced, more species will be observed. Thus, several extrapolation methods have been employed to infer the total richness (Chao and Chiu, 2016). The first is asymptotic approach based on species richness estimators. For example, some indicators infer the richness by taking into account the rare species present in the tail of the abundance distribution (Huggerth and Andersson, 2017), such as Chao1, ACE, and abundance-Jack1 (Abundance-based Coverage Estimator) (Heltshe and Forrester, 1983; Chazdon et al., 1998; Chao et al., 2014a), which could be classed into individual-based assessment. Alternatively, neglecting the species abundances, sample-based assessment only considers species accumulation with the increase in number of samples, with the representative indicators including Chao2, ICE, and incidence-Jack1 (Incidence-based Coverage Estimator) (Chao, 1987; Lee and Chao, 1994). The second is non-asymptotic approach based on rarefaction and extrapolation. The sample-size-based rarefaction and extrapolation of species richness has been proposed and recommended, which can be rarefied to small sample sizes or extrapolated to large sample sizes (Chao and Jost, 2012; Colwell et al., 2012). These types of asymptotic strategies construct a unified sampling framework for quantifying and comparing richness across multiple communities based on finite samples (Chao and Chiu, 2016). Overall, both non-asymptotic extrapolation and non-parametric asymptotic approaches can complement each other and provide a more accurate

estimation of species richness.

In another aspect, the indexes of α -diversity, i.e., Shannon entropy (Shannon, 1948) and inverse Simpson index (Simpson, 1949), have been regarded as the best choice to quantify abundance-based species diversity in both macro- and micro-ecological surveys (Haegeman et al., 2013; Alberdi and Gilbert, 2019). Recently, Chao et al. unified all commonly used α -diversity indexes as Hill numbers by setting up a weighted value (q) to the abundances (Ellison, 2010; Chao et al., 2014b). This unified statistical approach on measuring biodiversity has been refined and improved in recent years, but few studies have incorporated their diversity surveys through this approach (Haegeman et al., 2013; Kang et al., 2016; Hu et al., 2019). Since Hill number provides a convenient way to compare diversity estimates between different studies, it has been recommended for DNA based diversity analysis (Alberdi and Gilbert, 2019). Species richness, as a fundamental component of α -diversity, can be integrated to Hill numbers with the weighted value $q = 0$. As q rises, Hill number begins to take account of abundance and is increasingly influenced by high-abundance species, some other commonly used indices, including Shannon entropy ($q = 1$) and inverse Simpson index ($q = 2$), can be obtained to offer insights into dominance and other community characteristics.

In this study, we conducted an intensive sampling campaign (564 replicates in total) in a grassland in northern China by using a nested sampling design in four sampling quadrats (1 m² of topsoil). We attempted to address two fundamental questions through this deep sequencing. (i) How many prokaryotic (including bacterial and archaeal) species inhabit 1 m² of grassland topsoil (0–20 cm)? (ii) What are the impacts of sequencing depth on the α -diversity profiles through Hill number measurement?

2 Material and methods

2.1 Study site

Our study was conducted in a semiarid temperate steppe in Duolun County (42°02' N, 116°17' E, 1324 m a.s.l.), Inner Mongolia Autonomous Region of China. This grassland is located in a natural grazing region without fertilizer or pesticide usages. This site belongs to typical temperate zone habitat, and characterized by a semiarid continental monsoon climate. Annual mean temperature is 2.1°C and annual mean precipitation is ~385.5 mm (Zhang et al., 2017a). Soil is Haplic Calcisols (FAO classification) or chestnut soil (Chinese classification) with a mean bulk density of ~1.31 g cm⁻³. Perennials are dominated species, including *Artemisia frigida*, *Agropyron cristatum*, *Cleistogenes squarrosa*, and *Stipa krylovii* (Ru et al., 2018).

2.2 Soil sampling

We selected four square plots, each with an area of 1 m² (1 m

$\times 1$ m), as quadrats (named Q1, Q2, Q3 and Q4) from the steppe within a 2000 m transect. In each quadrat, 33 soil cores/samples (0–20 cm depth and 4 cm in diameter) were collected by a nest design in August 2017 (Fig. 1). Soils were sieved through a 2 mm mesh to remove roots and rocks. Samples were immediately stored in iceboxes and transferred to the laboratory. Thereafter, some technical replicates, including extracting DNA before or after soil sample mixing and different sequencing barcodes on each sample, were made to further ensure the robustness of our results. In all, each quadrat included 33 samples with total of 141 replicates. A detailed description of replicate design is available in our previous research (Li et al., 2021), as well as given in the Supplementary Methods.

2.3 Measurement of soil properties

Subsamples from each soil sample were collected to measure the total nitrogen (TN), ammonium nitrogen ($\text{NH}_4^+\text{-N}$), nitrate nitrogen ($\text{NO}_3^-\text{-N}$), and total organic carbon (TOC) at the Institute of Soil Science, Chinese Academy of Sciences (Nanjing, China). TN was measured by potassium persulfate oxidation method, $\text{NH}_4^+\text{-N}$ and $\text{NO}_3^-\text{-N}$ were quantified with NH_4Cl extraction method, TOC was estimated with potassium dichromate oxidation-ferric salt titration method (Bao, 2000). In addition, soil pH was measured in 1:2.5 (W/V) suspension of soil in distilled water. Soil moisture content was calculated by the weight loss after thoroughly dry.

2.4 DNA extraction, PCR and high-throughput sequencing

A total of 0.5 g soil was used to extract DNA via a FastDNA SPIN kit for soil (Qbiogene, Solon OH). The 16S rRNA gene was amplified using primer set 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3') targeting the V4 hypervariable region with sample-specific paired barcodes (Caporaso et al., 2012). The polymerase chain reaction (PCR) was performed in a 50 μL mixture, including 0.5 μL Taq DNA polymerase (TaKaRa), 5 μL of $10 \times$ PCR buffer, 1.5 μL dNTP mixture, 1.5 μL of 10 μM each primer, 1 μL of template DNA (20–30 ng μL^{-1}), and 39 μL ddH₂O. The reaction conditions were as follows: denaturation at 94°C for 1 min, 30 cycles of 94°C for 20 s, 57°C for 25 s, 68°C for 45 s, thereafter extension at 68°C for 10 min. PCR products were purified with E.Z.N.A.® Gel Extraction Kit (Omega Bio-Tek, Inc., USA), quantified by Nanodrop 2000 and pooled in equimolar concentrations. Sequencing library was constructed under the instructions of Illumina® VAHTSTM Nano DNA Library Prep Kit (Vazyme Biotech Co., Ltd). Sequencing was performed on Illumina HiSeq PE250 kit (Guangdong Magigene Biotechnology Co., Ltd).

2.5 Quantitative PCR (qPCR)

Extracted DNA was further quantified by qPCR to determine the 16S rRNA gene abundance. qPCR reaction was

performed in a 20 μL mixture with 10 μL SuperReal PreMix Plus (SYBR Green, TIANGEN, FP205), 10 ng of template DNA and 1.5 μL of 10 μM primer (10 μM ; Forward: 5'-CCTACGGGAGGCAGCAG-3'; Reverse: 5'-TTACCGCGGCTGCTGGCAC-3'). Each reaction was performed in triplicate with parallel negative controls under denaturing for 15 min at 95°C, 40 cycles of 10 s at 95°C, 30 s at 50°C and 32 s at 72°C in a CFX96 Touch Real-Time PCR Detection System (BioRad). Copy numbers of the 16S rRNA gene were normalized and calculated using the ΔCt method. According to previous studies, the number of 16S rRNA gene copies is approximately an order of magnitude greater than the number of total prokaryotes (Bressan et al., 2015; Zhang et al., 2017c). We used this knowledge to convert 16S rRNA gene copy number to prokaryotic cell number. Prokaryotic cell number was then applied to scaling laws (Locey and Lennon, 2016) to estimate species richness.

2.6 Data processing

Raw sequences were quality filtered and analyzed with a public Galaxy pipeline (<http://mem.rcees.ac.cn:8080>) (Feng et al., 2017). Sequences were demultiplexed according to sample-specific barcodes with one mismatch allowed, and then barcodes and primers were trimmed. Paired of sequences were merged by FLASH, and low-quality bases with an average quality score lower than 20 were discarded by Btrim (Kong, 2011; Magoc and Salzberg, 2011). Additionally, sequences with any ambiguous bases were removed and only sequences within 245–260 bp were kept. After quality trimming, Operational Taxonomic Units (OTUs) were generated using UPARSE at 97% sequencing similarity threshold (Edgar, 2013; Zhang et al., 2017b). Singletons, which are defined as OTUs with only one read across all replicates, were discarded. In addition, Deblur and DADA2 were also conducted to further explore their commonalities in microbial ecology studies and provide more messages in microbial richness estimation. Deblur was applied to generate Amplicon Sequence Variants (ASVs), only sequences which appear at least 10 times study wide and 1 time per sample were retained, other parameters were default options according to the recommendation from Deblur workflow (Amir et al., 2017). For DADA2 pipeline, we used standard filtering parameters in filterAndTrim function: maxN = 0 (sequences with ambiguous nucleotides), truncQ = 2 (truncate reads at the first instance of a quality score less than or equal to 2) and maxEE = 2 (maximum number of expected errors allowed in a read). Considering our big data set, “pseudo” option was used in dada function to perform sample inference. For other steps in DADA2, the workflow recommended by the DADA2 pipeline tutorial (1.8) was utilized to generate an ASV table followed the default parameters (Callahan et al., 2016).

2.7 Statistical analysis

For each different classification algorithms, data from 141

replicates in each quadrat were pooled before further analysis. Two approaches, including an asymptotic approach via species richness estimation, and a non-asymptotic approach via standardized sample size were applied to infer species richness. For asymptotic approach, various non-parametric richness estimators were used for estimating the theoretical number of species by fossil package in R project (Vavrek, 2011). To be specific, as estimates of unseen richness are based on observed rare species, either abundance (Chao1, ACE and Jack1) or incidence (Chao2, ICE and Jack1) based methods were performed for estimation, to reduce the preference of investigator and for mutual corroboration (Chao et al., 2005; Rajakaruna et al., 2016). For non-asymptotic approach, rarefaction and extrapolation curves were created on the basis of standardized sample size, to test whether the retained reads were enough to cover the vast majority of species and estimate the theoretical richness. Additionally, diversities based on Hill number were calculated. Non-asymptotic estimate approach and Hill numbers were implemented by iNEXT package in R project (Hsieh et al., 2016). Representative sequences obtained by UPARSE, Deblur and DADA2 were classified into different taxonomy by Bayesian classifier against the RDP training set, respectively (Wang et al., 2007). All results were visualized using ggplot2 package in R project (Ginestet, 2011).

2.8 Measurement of Hill numbers

Hill number (or effective number of species) means the number of equally abundant species that are required to give the same value of a diversity measure (Chao et al., 2014b). It's a mathematically unified parametric family of diversity indexes distinguishing themselves by a parameter (q), which determines the sensitivity of the index to species relative abundances. The Hill number of order q in a sample is

defined as (Hill, 1973):

$${}^qD = \left(\sum_{i=1}^S p_i^q \right)^{1/(1-q)}, \quad q \geq 0 \text{ and } q \neq 1$$

S is the species number, p_i is the relative abundance of species i . D is the diversity value, which includes the three most important diversity measures: species richness ($q = 0$), the exponential of Shannon entropy ($q = 1$), and the inverse of Simpson index ($q = 2$).

2.9 Data availability

Raw data after sequencing have been deposited in the Genome Sequence Archive (Wang et al., 2017) in BIG Data Center (Zhang et al., 2019a), Beijing Institute of Genomics (BIG), Chinese Academy of Sciences (<http://bigd.big.ac.cn/gsa>). Accession number is CRA001897.

3 Results

3.1 Sampling design, overall sequencing effort and community composition

To address accurate microbial diversity in 1 m² of topsoil, an intensive sampling effort was implemented across four quadrats positioned in a typical grassland (Fig. 1). Measurement of soil properties of these samples showed that there was a low heterogeneity within each quadrat, while there were some degrees of environmental heterogeneity among quadrats (Table S1). After quality control of the sequences, a total of 54 122 060 reads were retained, resulting in 16 145 431, 13 554 635, 11 659 856, and 12 762 138 reads for quadrats Q1, Q2, Q3, and Q4, respectively. These sequences were used for subsequent species classifications by UPARSE or

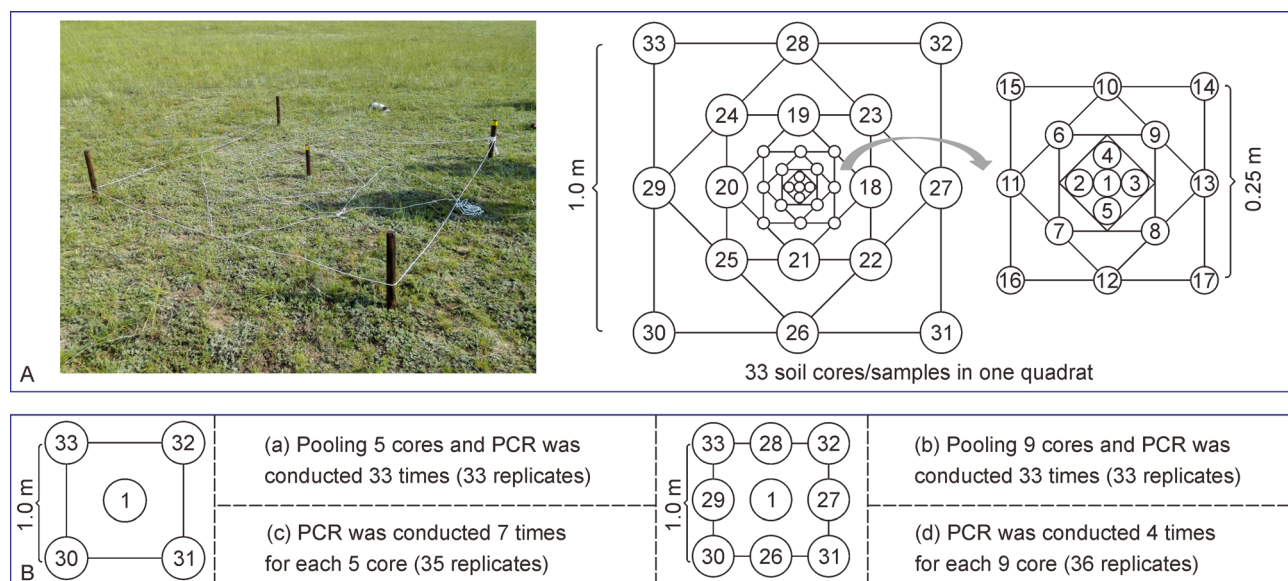


Fig. 1 Detailed experiment design for sampling in one quadrat. (A) 33 soil cores/samples in one quadrat. (B) Four pooling strategies.

Deblur. Meanwhile, DADA2 was conducted by raw sequencing output. After OTU or ASV generation, the four quadrats, respectively, retained 7 681 474 to 10 993 500 sequences by UPARSE, 3 733 659 to 5 194 295 sequences by Deblur and 7 697 276 to 10 815 483 sequences by DADA2 (Table S2).

The microbial communities across these four different quadrats were identified. 57.70%, 59.60% and 45.70% species were shared among all four quadrats for UPARSE, Deblur and DADA2, respectively (Supplemental Fig. S1). The number of species unique to each quadrat was very small in UPARSE (2.00% to 5.01%) and Deblur (0.55% to 3.49%), while this number was higher in DADA2 (8.24% to 12.8%). Among three classification algorithms, DADA2 identified the fewest phyla (34) and they were all within UPARSE's radius (39). Phyla obtained by both of DADA2 and UPARSE were included in Deblur (43). Highly similar community composition was observed via the different classification methods. Actinobacteria, Proteobacteria, Acidobacteria, Thaumarchaeota, Firmicutes, Verrucomicrobia, Bacteroidetes, Gemmatimonadetes, and Planctomycetes were the dominant phyla (abundance > 2%) present in grassland (Supplemental Fig. S2a). Abundance of Thaumarchaeota was higher by UPARSE, and DADA2 identified more abundance of Actinobacteria, Acidobacteria and Gemmatimonadetes. Additionally, genera belonged to Nitrososphaera, Gp4, Spartobacteria genera incertae sedis, Bacillus, Rubrobacter, Gp6, Gaiella, and Gemmatimonas were predominant (Supplemental Fig. S2b). UPARSE reckoned more abundance of Nitrososphaera than Deblur and DADA2.

3.2 The richness estimation by non-asymptotic extrapolation approach

In this study, the first crucial question we focused on was how many prokaryotic species inhabit 1 m² of topsoil. For the non-asymptotic approach, rarefaction and extrapolation curves were created based on standardized sample size (Fig. 2). Due to the extremely high sequencing depth, the rarefaction curve passed the inflection point and tended to flatten out. At this point, different classification algorithms could greatly impact the final number of OTUs or ASVs, both observed and estimated. The average observed richness was 24 921 ± 634 for UPARSE, 52 578 ± 1919 for Deblur and 15 699 ± 610 for DADA2, a 3.35-fold difference between the maximum and the minimum. To estimate the theoretical richness, the rarefaction curve was first extrapolated to double the sample size by using the default setting of iNEXT package in R (Fig. 2). The extrapolation curve was basically flat and estimated richness for UPARSE, Deblur and DADA2 were 26 655 ± 573, 55 844 ± 1373 and 15 807 ± 612, respectively. On that basis, if one continues to extrapolate until 30 or 20 million of reads, the number of species will increase slightly to 27 137 ± 487, 56 527 ± 1170 and 15 809 ± 613 for UPARSE, Deblur and DADA2 respectively. The following increases of richness could be negligible because all curves were approximately saturated (Fig. 2).

We may assume for the moment that the value for richness extrapolated to the maximum sample size is the value of theoretical richness, i.e., 27 137 for UPARSE, 56 527 for Deblur and 15 809 for DADA2. Based on this hypothesis, we examined the number of reads required to achieve different proportions of the theoretical richness (Table 1). Roughly 45 538 ± 6226, 81 235 ± 10 224 and 11 157 ± 829 reads for UPARSE, Deblur and DADA2, respectively, would only account for 20% of species in 1 m² grassland topsoil. For UPARSE and Deblur, nearly 500 000 sequences were required to reach 50% of species, while roughly 100 000 sequences were enough for DADA2. This number is much higher than the reads per sample, or even per site, currently obtained from high-throughput microbial sequencing studies, especially considering this number only applies to a small area (1 m²). To catch 80% richness, about 3 461 562 ± 297 567, 1 878 148 ± 174 972 and 911 446 ± 212 892 sequences were needed for UPARSE, Deblur and DADA2, respectively, and the numbers of sequences would be nearly twice on this basis to cover 90% richness. Furthermore, UPARSE could cover more diverse species with a small sample size, while Deblur captured more species when sample size was large, that is, Deblur requires less sequencing depth when the goal is to observed over 50% species in 1 m² grassland topsoil compared with UPARSE. Overall, DADA2 was the best at catching species under the same sequencing depth.

3.3 The richness estimation by non-parametric asymptotic approach

For asymptotic approach, six different non-parametric richness estimators, including Chao1, Chao2, ACE, ICE, abundance-Jack1, and incidence-Jack1, were chosen to estimate the theoretical species numbers in each quadrat. Richness estimations were 27 193 ± 1076 for UPARSE, 56 985 ± 2347 for Deblur, and 18 215 ± 3789 for DADA2 (Table 2). The estimated richness increased rapidly with the sample size at the beginning, and then gradually leveled off, while the ratio of estimated richness to observed richness declined quickly at first and tended to flatten off to approximately 1 at greater sequencing depth, except for three incidence-based coverage estimators (ICE, incidence-Jack1 and Chao2) by DADA2 (Fig. 3). We use Chao2 as a typical example to illustrate this result. According to the classical define of Chao2: $S_{\text{Chao2}} =$

$S_{\text{obs}} + \left(\frac{m-1}{m} \right) \frac{Q_1^2}{2Q_2}$, where S_{obs} is observed richness, m is sample number, Q_1 is the frequency of uniques (species that occur in only one sample) and Q_2 is the frequency of duplicates (species that occur in only two samples). This anomaly was mainly caused by abnormally increasing Q_1 with more sample size by DADA2 (Supplemental Fig. S3). Meanwhile, ICE and incidence-Jack1 also include Q_1 in their formular, so the effect was also visible. This anomaly suggested that DADA2 may be inapplicable to incidence-based coverage estimator in our study. Based on these results, we did not consider incidence-based coverage

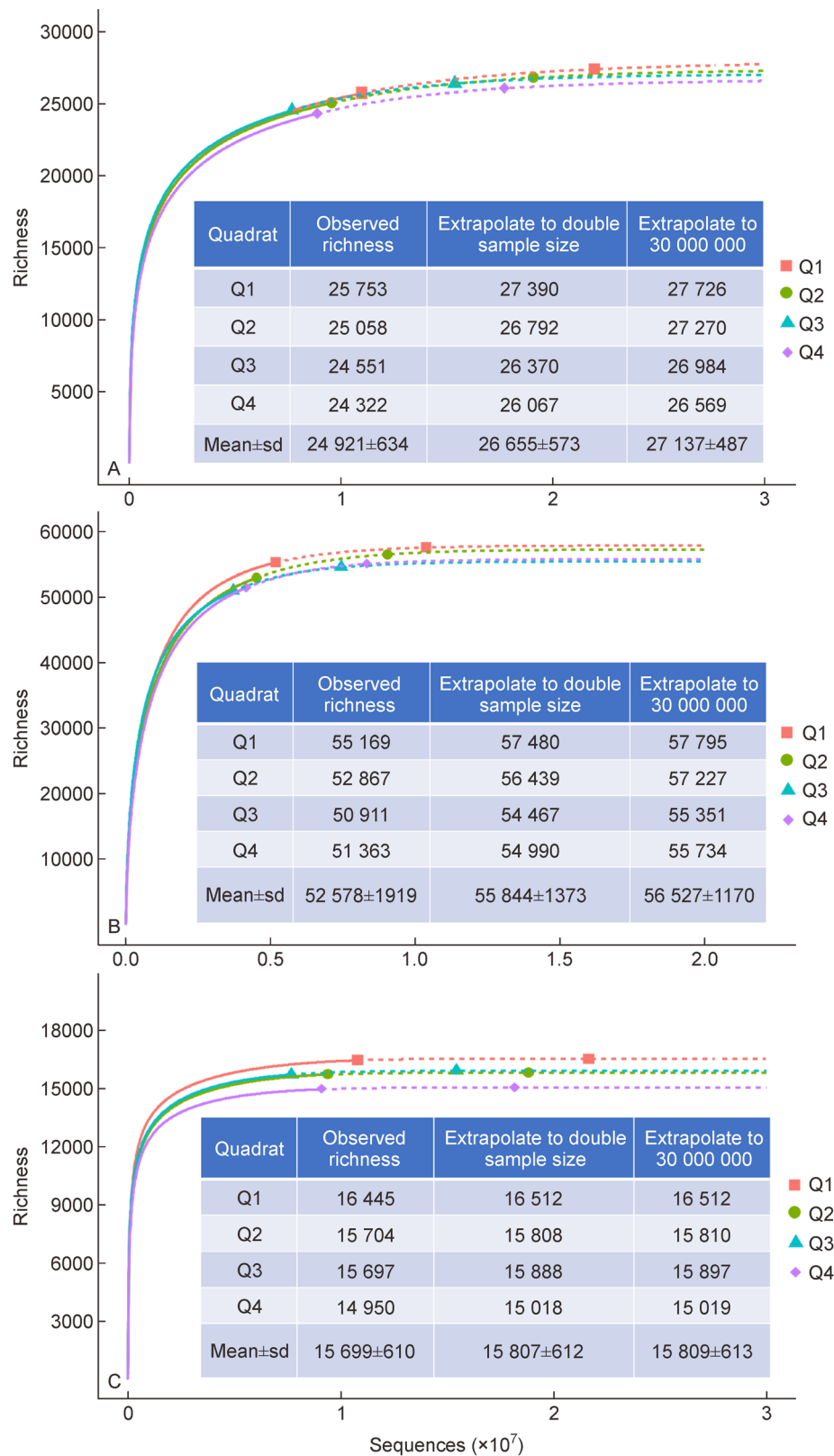
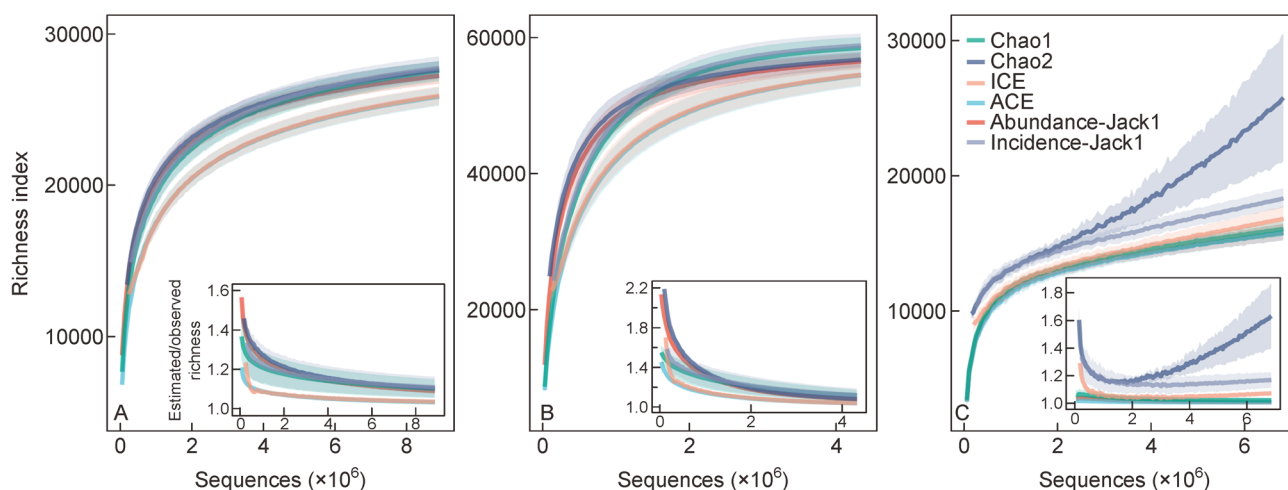


Fig. 2 Rarefaction and extrapolation curves by (A) UPARSE, (B) Deblur, and (C) DADA2. The points at the end of the solid line are the number of observed species. The points on the dotted line are the species number obtained by extrapolating to the double current sequence number. Embedded table shows observed or estimated richness under different sequences. sd is standard deviation.

Table 1 Required sequences to reach specific proportion of theoretical species.

Proportion of theoretical species	Sequences (UPARSE)	Sequences (Deblur)	Sequences (DADA2)
10%	18 616 \pm 2776	26 551 \pm 3720	3318 \pm 384
20%	45 538 \pm 6226	81 235 \pm 10 224	11 157 \pm 829
30%	110 625 \pm 13 428	166 617 \pm 17 656	25 729 \pm 2687
40%	253 096 \pm 16 536	293 464 \pm 29 653	52 563 \pm 6460
50%	501 266 \pm 42 870	481 417 \pm 49 278	105 194 \pm 18 894
60%	960 509 \pm 84 623	755 494 \pm 68 604	201 101 \pm 28 049
70%	1 817 972 \pm 157 057	1 178 895 \pm 103 394	405 203 \pm 69 567
80%	3 461 562 \pm 297 567	1 878 148 \pm 174 972	911 446 \pm 212 892
90%	7 015 759 \pm 688 290	3 277 020 \pm 383 091	2 456 694 \pm 840 372

**Fig. 3** Relationships of sequences and richness index (A) UPARSE, (B) Deblur, and (C) DADA2. The image embedded in the lower right corner is the ratio of estimated richness divided by observed richness with different sample size. The shaded regions around the curves are the average value \pm 95% confidence interval.

estimators for DADA2, and adjusted richness estimations were 15 919 \pm 193 (Adjusted value in Table 2). By comparing all estimators, Chao1 and Chao2 indexes were closest to the observed richness at all sequencing depths, while ACE (ICE) and abundance-Jack1 (incidence-Jack1) intersected as the sequencing deepened by UPARSE and Deblur.

While these non-parametric richness estimators have been widely applied in estimating microbial community richness, high number of reads, as seen in this study, are rarely obtained as the typical sampling effort generally returns 10 000 to 100 000 sequences per sample. In this context, we applied the estimated richness for each estimator in Table 2 as the number of theoretical richness. According to this result, we estimated the proportion of estimated richness to theoretical richness using 50 000 to 1 000 000 reads (Tables S3, S4, S5). Using 50 000 reads per sample, the range of estimates were 26.76% to 33.38% for UPARSE, 23.13% to 35.08% for Deblur, and 36.79% to 42.27% for DADA2. With a sampling effort of 1 000 000 reads, the highest richness estimation for UPARSE,

Deblur and DADA2 reached 75.44% from Chao1, 88.39% from Chao2 and 87.59% from Jack1 of the theoretical richness, respectively. These results highlight how insufficient sequencing depth will cause great deviations not only in observed richness, but also in estimated richness by commonly used richness estimation indexes.

3.4 Diversity characterized by high order of Hill numbers

Richness has become one of the most fundamental and important indicators of community diversity. However, this measurement does not consider the abundance of individuals within a community, limiting our understanding of α -diversity. Recently, richness (Hill, $q = 0$) has been unified into a general framework of Hill numbers. As all Hill numbers are presented as intuitive units of effective numbers of species, they can be directly compared across orders q , especially for Shannon entropy (Hill, $q = 1$) and inverse Simpson index (Hill, $q = 2$). Here we explored the relationships of sequencing depth on

Table 2 Observed and estimated richness by non-parametric asymptotic approach.

Methods	Observed richness	Chao1	Chao2	ACE	ICE	Abundance-Jack1	Incidence-Jack1	Average
UPARSE	24 921±634	27 255±539 (91.43±0.60%)	27 634±546 (90.18±0.58%)	25 855±611 (96.38±0.18%)	25 934±614 (96.09±0.17%)	28 113±585 (88.64±0.42%)	28 365±594 (87.85±0.40%)	27 193±1076
Deblur	52 578±1919	56 534±1173 (92.98±1.65%)	56 780±1127 (92.58±1.74%)	54 477±1618 (96.50±0.73%)	54 615±1618 (96.26±0.76%)	59 541±1188 (88.29±1.76%)	59 962±1198 (87.67±1.80%)	56 985±2347
DADA2	15 699±610	15 809±613 (99.30±0.39%)	25 764±4716 (62.05±8.16%)	15 806±616 (99.33±0.22%)	16 848±869 (93.23±1.33%)	16 141±635 (97.27±0.78%)	18 920±1047 (83.04±1.55%)	18 215±3879 (Adjusted: 15 919±193)

The values in parentheses are the ratio of the observed richness to the estimated richness. Adjusted value for DADA2 was the average value of three abundance-based coverage estimators (Chao1, ACE and Abundance-Jack1).

observed Hill numbers and estimated theoretical Hill numbers (Table 3). From our results, observed Hill numbers were almost the same as the estimated numbers. Estimated Shannon entropy was found to be among 1134 and 1915, and estimated inverse Simpson index were among 181 and 447 for three clustering methods. Only a small number of reads are required to calculate Hill values. For example, 10 000 reads were adequate to characterize over 50% and 95% of estimated Shannon entropy and inverse Simpson index within 1 m² of topsoil. Furthermore, we plotted observed Hill numbers profile as a continuous function of the parameter q (Supplemental Fig. S4). The steep declines of curves with a higher q demonstrated an uneven distribution of species relative abundances. Additionally, this diversity profile characterized the degree of dominance in the assemblage vividly. Sequencing depth mainly contributed in low q , and the curves of different sequencing depth, from 10 000 to all reads, gradually converged, especially with $q \geq 1$.

4 Discussion

4.1 The diversity profile of prokaryotes in 1 m² topsoil (0–20 cm depth) of grassland

Soil microbial communities are among the most important, diverse, and complex assemblages in the biosphere (Zhou et al., 2011). The richness of soil microorganisms has been studied for over a century. Although recent advances in marker gene sequencing technology and bioinformatic analysis have provided drastic improvements, the majority of prokaryotic diversity still remains undiscovered (Delgado-Baquerizo et al., 2018; Knight et al., 2018). The current consensus is that richness based on sampling data through marker gene sequencing technology is highly dependent on sample size (Colwell and Coddington, 1994; Chiu and Chao,

2016). In our study, a total of 141 replicates per a 1 m² quadrat were collected and sequenced, and by far the deepest sequencing of a small-scale area were obtained (Table S2). Additionally, although the larger spatial distance among the four quadrats lead to the existence of environmental heterogeneity, the community composition and the final observed richness were relatively stable (Supplemental FigureS1, S2). These provided a solid foundation to obtain accurate richness estimations. Taking together, the results of both the non-asymptotic extrapolation and non-parametric asymptotic approaches showed similar species richness (Fig. 3, Table 2). Measurements from UPARSE 97% similarity cutoff, Deblur and DADA2 can provide benchmarks for further diversity range between 27 137 \pm 487 to 27 193 \pm 1076, 56 527 \pm 1170 to 56 985 \pm 2347, and 15 809 \pm 613 to 15 919 \pm 193, respectively.

Richness is a central concept in understanding microbial communities. Over the past decade, numerous studies have been performed to try to uncover this fundamental measurement, especially for a gram of soil. Previous works have identified richness estimations of approximately 10 000 from the top 10 cm of soil from a boreal forest (Torsvik et al., 1990), 1 million for soil (Gans et al., 2005) while this number has been questioned (Bunge et al., 2006; Volkov et al., 2006), 2000 to 5000 for soil (Schloss and Handelsman, 2006), and 26 140 to 53 533 from the top 10 cm of forest soil (Roesch et al., 2007). Overall, estimates of the bacterial richness per gram of soil have varied between thousands to millions, greatly misleading ecological surveys of microorganisms. The reasons for this huge difference could be assigned to insufficient sequencing depth compared to the research scale, different estimation methods, and large biases hidden within some methods. Therefore, it is necessary to conduct estimates by using multiple feasible estimation methods with deep sequencing at appropriate scales. Recently, Locey and Lennon formulated a microbial richness-abundance scaling

Table 3 Relationships of sequencing depth to Hill numbers and coverages.

Reads ($\times 10^4$)	Methods					
	UPARSE		Deblur		DADA2	
	Shannon entropy (Hill, $q = 1$)	Inverse Simpson (Hill, $q = 2$)	Shannon entropy (Hill, $q = 1$)	Inverse Simpson (Hill, $q = 2$)	Shannon entropy (Hill, $q = 1$)	Inverse Simpson index (Hill, $q = 2$)
1	801.87 \pm 94.61	210.66 \pm 61.07	899.30 \pm 127.42	177.72 \pm 53.89	1333.34 \pm 139.27	427.36 \pm 119.77
2	903.60 \pm 111.83	213.01 \pm 62.29	1059.81 \pm 161.88	179.40 \pm 54.76	1530.81 \pm 168.77	437.13 \pm 124.48
3	952.14 \pm 120.15	214.07 \pm 63.07	1145.91 \pm 181.21	179.97 \pm 55.06	1622.44 \pm 182.74	440.50 \pm 126.11
5	1001.97 \pm 128.69	214.38 \pm 63.01	1243.62 \pm 203.62	180.43 \pm 55.29	1713.32 \pm 196.55	443.23 \pm 127.44
10	1051.51 \pm 137.03	214.92 \pm 63.29	1354.54 \pm 229.22	180.77 \pm 55.47	1798.77 \pm 209.19	445.30 \pm 128.46
50	1109.98 \pm 146.25	215.32 \pm 63.50	1516.37 \pm 264.59	181.05 \pm 55.62	1887.91 \pm 221.64	446.97 \pm 129.28
100	1120.64 \pm 147.82	215.37 \pm 63.52	1551.69 \pm 271.28	181.08 \pm 55.63	1901.57 \pm 223.46	447.18 \pm 129.38
Total reads	1132.82 \pm 149.31	215.42 \pm 63.55	1586.53 \pm 275.84	181.11 \pm 55.65	1915.20 \pm 225.21	447.37 \pm 129.47
Estimated value	1134.63 \pm 149.76	215.42 \pm 63.55	1597.73 \pm 278.84	181.11 \pm 55.65	1915.20 \pm 225.21	447.37 \pm 129.47

relationship which could be applied to estimate species richness based on the total abundance of prokaryotes, and concluded there were 10^{11} – 10^{12} microbial species on earth (Locey and Lennon, 2016). According to their unified scaling law: $S = 7.6 \times N^{0.35}$ and inputting our measured cell numbers in 1 m² grassland topsoil (0–20 cm), it was predicted that average richness would reach 973 510 in our quadrats (Table S6). This number was roughly between 36, 17 and 61-fold higher than the richness estimated using our data by UPARSE, Deblur and DADA2, respectively. As their dominance scaling law has not been verified on such a small scale, the inconsistencies with our results indicate that there may be potential deviations in its applications, especially at local scales.

4.2 The Effect of sequencing depth on diversity estimation

Certainly, the depth of sequencing showed a significant impact on the observed richness, which further influenced estimated richness (Table 1, Tables S3, S4, S5). It should be noted that there are some biases raised by various non-parametric richness estimators, including Chao1, Chao2, ACE, ICE abundance-Jack1, and incidence-Jack1. It is no doubt that these indices have been widely applied in microbiome associated analyses to compensate for the potential under-sampled data and estimate theoretical richness (Caporaso et al., 2011; Tu et al., 2016; Zhou et al., 2016; Deng et al., 2018; Zhang et al., 2019b). However, while these estimators did show small differences, they also showed mutual corroboration in this study. Our results strongly suggested that insufficient sequencing depth would still underestimate the final richness even when using these estimators. For example, by non-asymptotic extrapolation approach, nearly 500 000 sequences only reach 50% of species. While by non-parametric asymptotic approach, estimated richness based on 500 000 sequences could only achieve around 57% to 64%, 65% to 77% and 50% to 77% for UPARSE, Deblur and DADA2, respectively. At present, the retained sequences in a large number of studies have not exceeded 500 000 for one site, much less for one sample. In this state, not only the observed richness, but also the estimated richness will be grossly underestimated by about a third of total richness. Considering how small the area (1 m²) we focused on, future research needs to continue to increase the sequencing depth for a more accurate richness estimation in soil microbial surveys.

Under-sampling is a major source of bias for the diversity assessment, and these biases are typically more severe for Hill numbers with low orders of q (Chao and Jost, 2015). Higher order of q gives more weight to high abundance of species, which would largely reduce the effect of the undetected species. During our own analyses, in contrast to richness, Hill numbers of higher order could reach saturation with relatively fewer sequences (Table 3). Our results provided explicit evidence to demonstrate that when the focus is on exploring different characteristics of the community, the

number of sequencing reads required varies greatly. Additionally, Hill numbers, especially for Shannon entropy and inverse Simpson index, have been recommended to quantify and compare microbial taxonomic diversity (Haegeman et al., 2013; Alberdi and Gilbert, 2019). Our strong results proved the robustness of Shannon entropy and inverse Simpson index with their good consistency under discriminatory sequencing depths (Supplemental Fig. S4). As Hill numbers have excellent estimation properties, this method could greatly advance future research.

4.3 The Effect of clustering strategies on diversity measurement

A key, albeit often overlooked, factor in determining diversity is related to the algorithms used to cluster sequences into OTUs or ASVs. Producing accurate diversity information from millions of reads is a primary requirement of all marker gene studies. To achieve this goal, many clustering methods based on different algorithms, have been implemented in recent years (Nguyen et al., 2015). However, inconsistency among algorithms is increased by the way they consider species with low abundances. UPARSE, a widely applied method, suggests removing singletons (Edgar, 2013), while Deblur and DADA2 use error profiles to obtain putative error-free reads at single-nucleotide resolution and further generate ASVs (Callahan et al., 2016; Amir et al., 2017). The recommended parameters for Deblur suggest keeping only the sequences which appeared a minimum of 10 times within a study. In our study, even though we followed all of the default recommendations, there were still large differences between the algorithms, e.g., richness derived from Deblur was nearly twice as much as UPARSE, and 3.3-fold more than DADA2. Regardless of the specific method chosen, it is difficult to robustly distinguish sequencing errors from true rare species. To guarantee the biological significance of the retained sequences, some low-abundance sequences must be discarded, which inevitably leads to the loss of richness. Indeed, previous studies have pointed out that rare species cause a serious problem for estimating species richness (Mao and Colwell, 2005; Haegeman et al., 2013). While there is no silver bullet, the choice will rely on the relative risk of pseudo richness versus taxon bias in addressing particular ecological questions. In our study, our focus is not to compare the difference among these methods and explore the behind mechanism, instead, we try to use various widely applied methods to provide as much messages as possible, for reference in future research.

5 Conclusion

It is important to note that quantifying the astonishing diversity of soil microorganism is a great challenge. Based on our analyses from non-asymptotic extrapolation and non-parametric asymptotic approaches at a small and controllable

scale by thorough sampling, we deduced that the number of prokaryotic species was on the order of $27\,137 \pm 487$ to $27\,193 \pm 1076$, $56\,527 \pm 1170$ to $56\,985 \pm 2347$, and $15\,809 \pm 613$ to $15\,919 \pm 193$ by UPARSE 97% cutoff, Deblur and DADA2 in 1 m^2 of grassland topsoil, respectively. We acknowledged that our estimation was not exactly accurate due to the biases inherent to the marker gene sequencing process (Gohl et al., 2016), technical errors such as potential arbitrariness of treating low abundance sequences in OTU or ASV clustering (Balint et al., 2016), and internal limitations of non-parametric or parametric estimators (O'Hara, 2005; Bunge et al., 2014). However, by using these massive sequences, our results still provided a valuable benchmark for future research. Given our results, further improvement of sequencing depth is needed to better capture undetected observed as well as estimated species richness. In addition, if a study focuses on the α -diversity as represented by higher order Hill numbers, the sequencing depth can be appropriately downregulated to lower the cost. Also, a deeper and more thorough view of species diversity at this small scale will contribute greatly to the extrapolation at larger scales in the future studies.

Conflict of interest

The authors declare that they have no conflict of interest.

Acknowledgments

This study was supported by the National Natural Science Foundation of China (NSFC Grant No. U1906223), the National Key Research and Development Program (Grant No. 2019YFC1905001). The authors are very grateful to Dr. James Walter Voordeckers for careful edition on the final version. We thank the convenience provided by Restoration Ecology Experimental Demonstration Research Station of Institute of Botany, CAS.

Electronic supplementary material

Supplementary material is available in the online version of this article at <https://doi.org/10.1007/s42832-021-0111-5> and is accessible for authorized users.

Reference

Alberdi, A., Gilbert, M.T.P., 2019. A guide to the application of Hill numbers to DNA-based diversity analyses. *Molecular Ecology Resources* 19, 804–817.

Amir, A., McDonald, D., Navas-Molina, J.A., Kopylova, E., Morton, J. T., Xu, Z.Z., Kightley, E.P., Thompson, L.R., Hyde, E.R., Gonzalez, A., Knight, R., 2017. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems* 2, e00191–e16.

Balint, M., Bahram, M., Eren, A.M., Faust, K., Fuhrman, J.A., Lindahl, B., O'Hara, R.B., Opik, M., Sogin, M.L., Unterseher, M., Tedersoo, L., 2016. Millions of reads, thousands of taxa: microbial community

structure and associations analyzed via marker genes. *FEMS Microbiology Reviews* 40, 686–700.

Bao, S., 2000. *Soil and Agricultural Chemistry Analysis*. China Agricultural Press, Beijing.

Bressan, M., Gattin, I.T., Desaire, S., Castel, L., Gangneux, C., Laval, K., 2015. A rapid flow cytometry method to assess bacterial abundance in agricultural soil. *Applied Soil Ecology* 88, 60–68.

Bunge, J., Epstein, S.S., Peterson, D.G., 2006. Comment on "Computational improvements reveal great bacterial diversity and high metal toxicity in soil". *Science* 313, 918c.

Bunge, J., Willis, A., Walsh, F., 2014. Estimating the number of species in microbial diversity studies. *Annual Review of Statistics and Its Application* 1, 427–445.

Callahan, B.J., McMurdie, P.J., Rosen, M.J., Han, A.W., Johnson, A.J. A., Holmes, S.P., 2016. DADA2: High-resolution sample inference from Illumina amplicon data. *Nature Methods* 13, 581–583.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Huntley, J., Fierer, N., Owens, S.M., Betley, J., Fraser, L., Bauer, M., Gormley, N., Gilbert, J.A., Smith, G., Knight, R., 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME Journal* 6, 1621–1624.

Caporaso, J.G., Lauber, C.L., Walters, W.A., Berg-Lyons, D., Lozupone, C.A., Turnbaugh, P.J., Fierer, N., Knight, R., 2011. Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America* 108, 4516–4522.

Chao, A., 1987. Estimating the population-size for capture recapture data with unequal catchability. *Biometrics* 43, 783–791.

Chao, A., Chazdon, R.L., Colwell, R.K., Shen, T.J., 2005. A new statistical approach for assessing similarity of species composition with incidence and abundance data. *Ecology Letters* 8, 148–159.

Chao, A., Chiu, C.H., 2016. Nonparametric Estimation and Comparison of Species Richness 1–11.

Chao, A., Gotelli, N.J., Hsieh, T.C., Sander, E.L., Ma, K.H., Colwell, R. K., Ellison, A.M., 2014a. Rarefaction and extrapolation with Hill numbers: a framework for sampling and estimation in species diversity studies. *Ecological Monographs* 84, 45–67.

Chao, A., Jost, L., 2012. Coverage-based rarefaction and extrapolation: standardizing samples by completeness rather than size. *Ecology* 93, 2533–2547.

Chao, A., Jost, L., 2015. Estimating diversity and entropy profiles via discovery rates of new species. *Methods in Ecology and Evolution* 6, 873–882.

Chao, A.N., Chiu, C.H., Jost, L., 2014b. Unifying species diversity, phylogenetic diversity, functional diversity, and related similarity and differentiation measures through Hill numbers. *Annual Review of Ecology, Evolution, and Systematics* 45, 297–324.

Chazdon, R.L., Colwell, R.K., Denslow, J.S., Guariguata, M.R., 1998. Statistical methods for estimating species richness of woody regeneration in primary and secondary rain forests of northeastern Costa Rica. *Forest Biodiversity Research. Monitoring and Modeling* 20, 285–309.

Chiu, C.H., Chao, A., 2016. Estimating and comparing microbial diversity in the presence of sequencing errors. *PeerJ* 4, e1634.

Colwell, R.K., Chao, A., Gotelli, N.J., Lin, S.Y., Mao, C.X., Chazdon, R. L., Longino, J.T., 2012. Models and estimators linking individual-

- based and sample-based rarefaction, extrapolation and comparison of assemblages. *Journal of Plant Ecology* 5, 3–21.
- Colwell, R.K., Coddington, J.A., 1994. Estimating terrestrial biodiversity through extrapolation. *Philosophical Transactions of the Royal Society of London. Series B, Biological Sciences* 345, 101–118.
- Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-Gonzalez, A., Eldridge, D.J., Bardgett, R.D., Maestre, F.T., Singh, B.K., Fierer, N., 2018. A global atlas of the dominant bacteria found in soil. *Science* 359, 320–325.
- Deng, Y., Ning, D.L., Qin, Y.J., Xue, K., Wu, L.Y., He, Z.L., Yin, H.Q., Liang, Y.T., Buzzard, V., Michaletz, S.T., Zhou, J.Z., 2018. Spatial scaling of forest soil microbial communities across a temperature gradient. *Environmental Microbiology* 20, 3504–3513.
- Edgar, R.C., 2013. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nature Methods* 10, 996–998.
- Ellison, A.M., 2010. Partitioning diversity. *Ecology* 91, 1962–1963.
- Feng, K., Zhang, Z.J., Cai, W.W., Liu, W.Z., Xu, M.Y., Yin, H.Q., Wang, A.J., He, Z.L., Deng, Y., 2017. Biodiversity and species competition regulate the resilience of microbial biofilm community. *Molecular Ecology* 26, 6170–6182.
- Gans, J., Wolinsky, M., Dunbar, J., 2005. Computational improvements reveal great bacterial diversity and high metal toxicity in soil. *Science* 309, 1387–1390.
- Ginestet, C., 2011. ggplot2: Elegant Graphics for Data Analysis. *Journal of the Royal Statistical Society Series a-Statistics in Society* 174, 245–245.
- Gohl, D.M., Vangay, P., Garbe, J., MacLean, A., Hauge, A., Becker, A., Gould, T.J., Clayton, J.B., Johnson, T.J., Hunter, R., Knights, D., Beckman, K.B., 2016. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nature Biotechnology* 34, 942–949.
- Gotelli, N.J., Colwell, R.K., 2001. Quantifying biodiversity: procedures and pitfalls in the measurement and comparison of species richness. *Ecology Letters* 4, 379–391.
- Haegeman, B., Hamelin, J., Moriarty, J., Neal, P., Dushoff, J., Weitz, J. S., 2013. Robust estimation of microbial diversity in theory and in practice. *ISME Journal* 7, 1092–1101.
- Heltsh, J.F., Forrester, N.E., 1983. Estimating species richness using the Jackknife procedure. *Biometrics* 39, 1–11.
- Hill, M.O., 1973. Diversity and evenness: A unifying notation and its consequences. *Ecology* 54, 427–432.
- Hsieh, T.C., Ma, K.H., Chao, A., 2016. iNEXT: an R package for rarefaction and extrapolation of species diversity (Hill numbers). *Methods in Ecology and Evolution* 7, 1451–1456.
- Hu, Y.J., Veresoglou, S.D., Tedersoo, L., Xu, T.L., Ge, T.D., Liu, L., Chen, Y.L., Hao, Z.P., Su, Y.R., Rillig, M.C., Chen, B.D., 2019. Contrasting latitudinal diversity and co-occurrence patterns of soil fungi and plants in forest ecosystems. *Soil Biology & Biochemistry* 131, 100–110.
- Hugerth, L.W., Andersson, A.F., 2017. Analysing microbial community composition through amplicon sequencing: From sampling to hypothesis testing. *Frontiers in Microbiology* 8, 1561.
- Kang, S., Rodrigues, J.L.M., Ng, J.P., Gentry, T.J., 2016. Hill number as a bacterial diversity measure framework with high-throughput sequence data. *Scientific Reports* 6, 38263.
- Knight, R., Vrbanac, A., Taylor, B.C., Aksenov, A., Callewaert, C., Debelius, J., Gonzalez, A., Kosciulek, T., McCall, L.I., McDonald, D., Melnik, A.V., Morton, J.T., Navas, J., Quinn, R.A., Sanders, J. G., Swafford, A.D., Thompson, L.R., Tripathi, A., Xu, Z.J.Z., Zaneveld, J.R., Zhu, Q.Y., Caporaso, J.G., Dorrestein, P.C., 2018. Best practices for analysing microbiomes. *Nature Reviews. Microbiology* 16, 410–422.
- Kong, Y., 2011. Btrim: A fast, lightweight adapter and quality trimming program for next-generation sequencing technologies. *Genomics* 98, 152–153.
- Lee, S.M., Chao, A., 1994. Estimating population-size via sample coverage for closed capture-recapture models. *Biometrics* 50, 88–97.
- Li, S., Deng, Y., Du, X., Feng, K., Wu, Y., He, Q., Wang, Z., Liu, Y., Wang, D., Peng, X., Zhang, Z., Escalas, A., Qu, Y., 2021. Sampling cores and sequencing depths affected the measurement of microbial diversity in soil quadrats. *Science of the Total Environment* 767, 144966.
- Locey, K.J., Lennon, J.T., 2016. Scaling laws predict global microbial diversity. *Proceedings of the National Academy of Sciences of the United States of America* 113, 5970–5975.
- Magoc, T., Salzberg, S.L., 2011. FLASH: fast length adjustment of short reads to improve genome assemblies. *Bioinformatics (Oxford, England)* 27, 2957–2963.
- Mao, C.X., Colwell, R.K., 2005. Estimation of species richness: Mixture models, the role of rare species, and inferential challenges. *Ecology* 86, 1143–1153.
- Nguyen, N.H., Smith, D., Peay, K., Kennedy, P., 2015. Parsing ecological signal from noise in next generation amplicon sequencing. *New Phytologist* 205, 1389–1393.
- O'Hara, R.B., 2005. Species richness estimators: how many species can dance on the head of a pin? *Journal of Animal Ecology* 74, 375–386.
- Rajakaruna, H., Drake, D.A.R., Chan, F.T., Bailey, S.A., 2016. Optimizing performance of nonparametric species richness estimators under constrained sampling. *Ecology and Evolution* 6, 7311–7322.
- Roesch, L.F., Fulthorpe, R.R., Riva, A., Casella, G., Hadwin, A.K.M., Kent, A.D., Daroub, S.H., Camargo, F.A.O., Farmerie, W.G., Triplett, E.W., 2007. Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME Journal* 1, 283–290.
- Ru, J.Y., Zhou, Y.Q., Hui, D.F., Zheng, M.M., Wan, S.Q., 2018. Shifts of growing-season precipitation peaks decrease soil respiration in a semiarid grassland. *Global Change Biology* 24, 1001–1011.
- Schloss, P.D., Handelsman, J., 2006. Toward a census of bacteria in soil. *PLoS Computational Biology* 2, 786–793.
- Shannon, C.E., 1948. A mathematical theory of communication. *Bell System Technical Journal* 27, 379–423.
- Simpson, E.H., 1949. Measurement of diversity. *Nature* 163, 688–688.
- Tedersoo, L., Bahram, M., Polme, S., Koljalg, U., Yorou, N.S., Wijesundera, R., Ruiz, L.V., Vasco-Palacios, A.M., Thu, P.Q., Suija, A., Smith, M.E., Sharp, C., Saluveer, E., Saitta, A., Rosas, M., Riit, T., Ratkowsky, D., Pritsch, K., Poldmaa, K., Piepenbring, M., Phosri, C., Peterson, M., Parts, K., Partel, K., Otsing, E., Nouhra, E., Njouonkou, A.L., Nilsson, R.H., Morgado, L.N., Mayor, J., May, T.W., Majuakim, L., Lodge, D.J., Lee, S.S., Larsson, K.H., Kohout, P., Hosaka, K., Hiiesalu, I., Henkel, T.W., Harend, H., Guo,

- L.D., Greslebin, A., Grelet, G., Geml, J., Gates, G., Dunstan, W., Dunk, C., Drenkhan, R., Dearnaley, J., De Kesel, A., Dang, T., Chen, X., Buegger, F., Brearley, F.Q., Bonito, G., Anslan, S., Abell, S., Abarenkov, K., 2014. Global diversity and geography of soil fungi. *Science* 346, 1078.
- Torsvik, V., Goksoyr, J., Daae, F.L., 1990. High diversity in DNA of soil bacteria. *Applied and Environmental Microbiology* 56, 782–787.
- Tu, Q.C., Deng, Y., Yan, Q.Y., Shen, L.N., Lin, L., He, Z.L., Wu, L.Y., Van Nostrand, J.D., Buzzard, V., Michaletz, S.T., Enquist, B.J., Weiser, M.D., Kaspari, M., Waide, R.B., Brown, J.H., Zhou, J.Z., 2016. Biogeographic patterns of soil diazotrophic communities across six forests in the North America. *Molecular Ecology* 25, 2937–2948.
- Vavrek, M.J., 2011. fossil: Palaeoecological and palaeogeographical analysis tools. *Palaeontologia Electronica* 14:1T
- Volkov, I., Banavar, J.R., Maritan, A., 2006. Comment on “Computational improvements reveal great bacterial diversity and high metal toxicity in soil”. *Science* 313, 918.
- Wang, Q., Garrity, G.M., Tiedje, J.M., Cole, J.R., 2007. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Applied and Environmental Microbiology* 73, 5261–5267.
- Wang, Y.Q., Song, F.H., Zhu, J.W., Zhang, S.S., Yang, Y.D., Chen, T. T., Tang, B.X., Dong, L.L., Ding, N., Zhang, Q., Bai, Z.X., Dong, X. N., Chen, H.X., Sun, M.Y., Zhai, S., Sun, Y.B., Yu, L., Lan, L., Xiao, J.F., Fang, X.D., Lei, H.X., Zhang, Z., Zhao, W.M., 2017. GSA: Genome Sequence Archive. *Genomics, Proteomics & Bioinformatics* 15, 14–18.
- Wu, L.W., Ning, D.L., Zhang, B., Li, Y., Zhang, P., Shan, X.Y., Zhang, Q.T., Brown, M., Li, Z.X., Van Nostrand, J.D., Ling, F.Q., Xiao, N.J., Zhang, Y., Vierheilig, J., Wells, G.F., Yang, Y.F., Deng, Y., Tu, Q.C., Wang, A.J., Zhang, T., He, Z.L., Keller, J., Nielsen, P.H., Alvarez, P. J.J., Criddle, C.S., Wagner, M., Tiedje, J.M., He, Q., Curtis, T.P., Stahl, D.A., Alvarez-Cohen, L., Rittmann, B.E., Wen, X.H., Zhou, J. Z., Acevedo, D., Agullo-Barcelo, M., Andersen, G.L., de Araujo, J. C., Boehnke, K., Bond, P., Bott, C.B., Bovio, P., Brewster, R.K., Bux, F., Cabezas, A., Cabrol, L., Chen, S., Etchebehere, C., Ford, A., Frigon, D., Gomez, J.S., Griffin, J.S., Gu, A.Z., Habagil, M., Hale, L., Hardeman, S.D., Harmon, M., Horn, H., Hu, Z.Q., Jauffur, S., Johnson, D.R., Keucken, A., Kumari, S., Leal, C.D., Lebrun, L. A., Lee, J., Lee, M., Lee, Z.M.P., Li, M.Y., Li, X., Liu, Y., Luthy, R.G., Mendonca-Hagler, L.C., de Menezes, F.G.R., Meyers, A.J., Mohebbi, A., Oehmen, A., Palmer, A., Parameswaran, P., Park, J., Patsch, D., Reginatto, V., de los Reyes, F.L., Robles, A.N., Rossetti, S., Sidhu, J., Sloan, W.T., Smith, K., de Sousa, O.V., Stephens, K., Tian, R.M., Tooker, N.B., Vasconcelos, D.D., Wakelin, S., Wang, B., Weaver, J.E., West, S., Wilmes, P., Woo, S.G., Wu, J.H., Wu, L.Y., Xi, C.W., Xu, M.Y., Yan, T., Yang, M., Young, M., Yue, H.W., Zhang, Q., Zhang, W., Zhang, Y., Zhou, H. D., Brown, M., Consortium, G.W.M., 2019. Global diversity and biogeography of bacterial communities in wastewater treatment plants. *Nature Microbiology* 4, 1183–1195.
- Zhang, X.M., Johnston, E.R., Li, L.H., Konstantinidis, K.T., Han, X.G., 2017a. Experimental warming reveals positive feedbacks to climate change in the Eurasian Steppe. *ISME Journal* 11, 885–895.
- Zhang, X.X., Zhang, R.J., Gao, J.S., Wang, X.C., Fan, F.L., Ma, X.T., Yin, H.Q., Zhang, C.W., Feng, K., Deng, Y., 2017b. Thirty-one years of rice-rice-green manure rotations shape the rhizosphere microbial community and enrich beneficial bacteria. *Soil Biology & Biochemistry* 104, 208–217.
- Zhang, Z., Qu, Y., Li, S., Feng, K., Wang, S., Cai, W., Liang, Y., Li, H., Xu, M., Yin, H., Deng, Y., 2017c. Soil bacterial quantification approaches coupling with relative abundances reflecting the changes of taxa. *Scientific Reports* 7, 4837.
- Zhang, Z., Zhao, W.M., Xiao, J.F., Bao, Y.M., Wang, F., Hao, L.L., Zhu, J.W., Chen, T.T., Zhang, S.S., Chen, X., Tang, B.X., Zhou, Q., Wang, Z.H., Dong, L.L., Wang, Y.Q., Ma, Y.K., Zhang, Z.W., Wang, Z., Chen, M.L., Tian, D.M., Li, C.P., Teng, X.F., Du, Z.L., Yuan, N., Zeng, J.Y., Wang, J.Y., Shi, S., Zhang, Y.D., Wang, Q., Pan, M.Y., Qian, Q.H., Song, S.H., Niu, G.Y., Li, M., Xia, L., Zou, D., Zhang, Y. S., Sang, J., Li, M.W., Zhang, Y., Wang, P., Gao, Q.W., Liang, F., Li, R.J., Liu, L., Cao, J., Abbasi, A.A., Shireen, H., Li, Z., Xiong, Z., Jiang, M.Y., Guo, T.K., Li, Z.H., Zhang, H., Ma, L., Gao, R., Zhang, T., Li, W.L., Zhang, X.Q., Lan, L., Zhai, S., Zhang, Y.P., Wang, G.D., Wang, Z.N., Xue, Y.B., Sun, Y.B., Yu, L., Sun, M.Y., Chen, H.X., Hu, H., Guo, A.Y., Lin, S.F., Xue, Y., Wang, C.W., Ning, W.S., Zhang, Y., Luo, H., Gao, F., Guo, Y.P., Zhang, Q., Zhou, J.Q., Huang, Z., Cui, Q.H., Miao, Y.R., Ruan, C., Yuan, C.H., Chen, M., Jinpu, J., Gao, G., Xu, H.D., Li, Y.M., Li, C.Y., Tang, Q., Peng, D., Deng, W.K., Members, B.D.C., 2019a. Database resources of the BIG Data Center in 2019. *Nucleic Acids Research* 47, D8–D14.
- Zhang, Z.J., Deng, Y., Feng, K., Cai, W.W., Li, S.Z., Yin, H.Q., Xu, M. Y., Ning, D.L., Qu, Y.Y., 2019b. Deterministic assembly and diversity gradient altered the biofilm community performances of bioreactors. *Environmental Science & Technology* 53, 1315–1324.
- Zhou, J.Z., Deng, Y., Shen, L.N., Wen, C.Q., Yan, Q.Y., Ning, D.L., Qin, Y.J., Xue, K., Wu, L.Y., He, Z.L., Voordeckers, J.W., Van Nostrand, J.D., Buzzard, V., Michaletz, S.T., Enquist, B.J., Weiser, M.D., Kaspari, M., Waide, R., Yang, Y.F., Brown, J.H., 2016. Temperature mediates continental-scale diversity of microbes in forest soils. *Nature Communications* 7, 12083.
- Zhou, J.Z., Wu, L.Y., Deng, Y., Zhi, X.Y., Jiang, Y.H., Tu, Q.C., Xie, J.P., Van Nostrand, J.D., He, Z.L., Yang, Y.F., 2011. Reproducibility and quantitation of amplicon sequencing-based detection. *ISME Journal* 5, 1303–1313.