Assessment of microbial α-diversity in one meter squared topsoil

Shuzhen Li[a,b], Xiongfeng Du[a,c], Kai Feng[a,c], Yueni Wu[a,c], Qing He[a,c], Zhujun Wang[a,c], Yangying Liu[a,c], Danrui Wang[a,c], Xi Peng[a,c], Zhaojing Zhang[d], Arthur Escalas[e], Yuanyuan Qu[b], Ye Deng[a,c,d*]

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

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

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

d. Institute for Marine Science and Technology, Shandong University, Qingdao, China

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

* Corresponding author:
Ye Deng, E-mail: yedeng@rcees.ac.cn

Running title: Microbial α-diversity profile

Supplementary Methods

Replicate design

Detailed replicate information in one quadrat was summarized graphically in Fig. 1. Replicates could be divided into two parts, including the original soil cores samples and the generated replicates based on these cores. In each quadrat, 33 original soil cores samples with nested distribution were taken as biological replicates. The generated replicates could be constructed in four different soil pooling methods, which contained (a) 5 pooled biological replicates (the 1st, 30th, 31st, 32nd, 33rd in all 33 cores) with 33 technical replicates, and (b) 9 pooled biological replicates (the 1st, 26th, 27th, 28th, 29th, 30th, 31st, 32nd, 33rd in all 33 cores) with 33 technical replicates. These two treatments referred to pooling an equal amount of soil from these 5 or 9 cores, respectively, and PCR was conducted 33 times on each pool of cores, yielding 33 technical replicates. Additionally, there were (c) 5 biological replicates (the 1st, 30th, 31st, 32nd, 33rd in all 33 cores) each technically replicated 7 times and (d) 9 biological replicates (the 1st, 30th, 31st, 32nd, 33rd in all 33 cores) each technically replicated 4 times, and the difference was the relative number of biological replicates versus technical replicates. For (c), PCR was conducted 7 times on 5 biological replicates, yielding 5 biological × 7 technical = 35 replicates. For (d), PCR was conducted 4 times on 9 biological replicates, yielding 9 biological × 4 technical = 36 replicates. In addition to the above four pooling soil strategies, (e) data from all 33 cores were pooled in data analysis stage and was regarded as pooling data strategy. At the stage of our experimental design, we designed a similar number of replicate sizes for each treatment, *i.e.*, 33 for (a, b), 35 for (c), 36 for (d) and 33 for (e), to ensure that the different mixing schemes could achieve a similar sequencing depth.

Overall, each quadrat had 33+33+35+36+33 = 170 replicates in total, including 33 original samples and 137 generated replicates. In order to simplify the experimental replicates size during our experiments, some replicates could be replaced by others. Specifically, replicates with (c) from cores 1, 30, 31, 32, 33 with 4 times PCR could be substituted by the corresponding replicates with (d). For example, (c)-1-1, (c)-1-2, (c)-1-3, (c)-1-4 were four replicates from core 1 with 4 times PCR in (d), and these four replicates could also be regarded as (d)-1-1, (d)-1-2, (d)-1-3, (d)-1-4. We could reduce the replicate size by 20 in this manner. Additionally, original soil core sample from cores 1, 26, 27, 28, 29, 30, 31, 32, 33 could also be replaced by corresponding replicates with replicate in (c). Replicate size could be reduced by a further 9 in this manner. Hence, a total of 29 replicates could be integrated, final number of replicates was 141 in each quadrat and total replicates number was 564, covering both the original and the generated replicates.

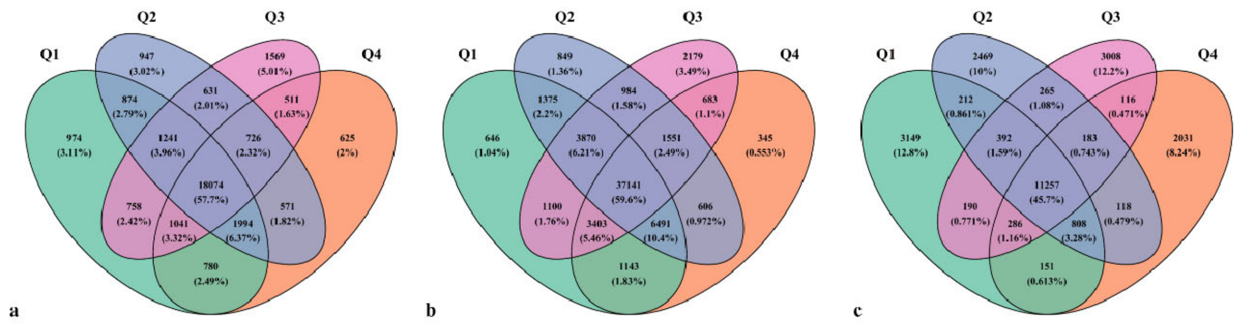

Figure S1 Venn diagrams of four quadrats (Q1, Q2, Q3 and Q4) by different algorithms. (a) UPARSE; (b) Deblur; and (c) DADA2.

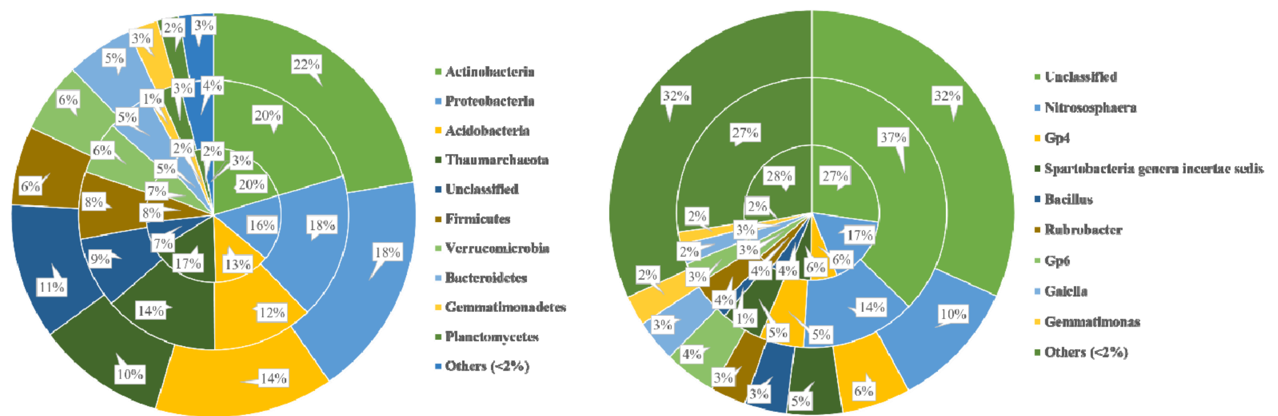

Figure S2 Taxonomic diversity at the (a) phylum and (b) genus levels. Percentages are average values of four quadrats (Q1, Q2, Q3 and Q4). Inner pie, UPARSE; middle pie, Deblur; outer pie, DADA2.

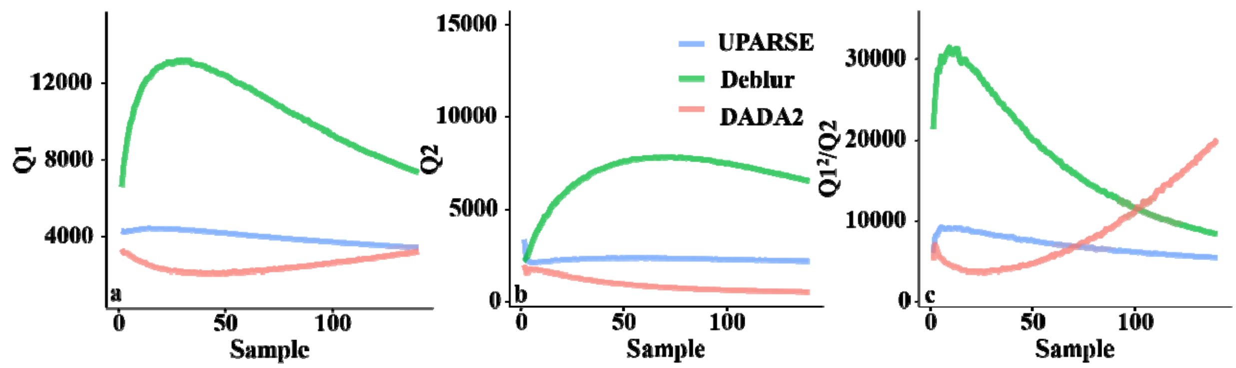

Fig. S3 (a), Q_1 (number of species that occur in exactly 1 samples); (b), Q_2 (number of species that occur in exactly 2 samples); and (c), Q_1^2/Q_2 of UPARSE, Deblur and DADA2. The shaded regions around the curves are the average value ± 95% confidence interval.

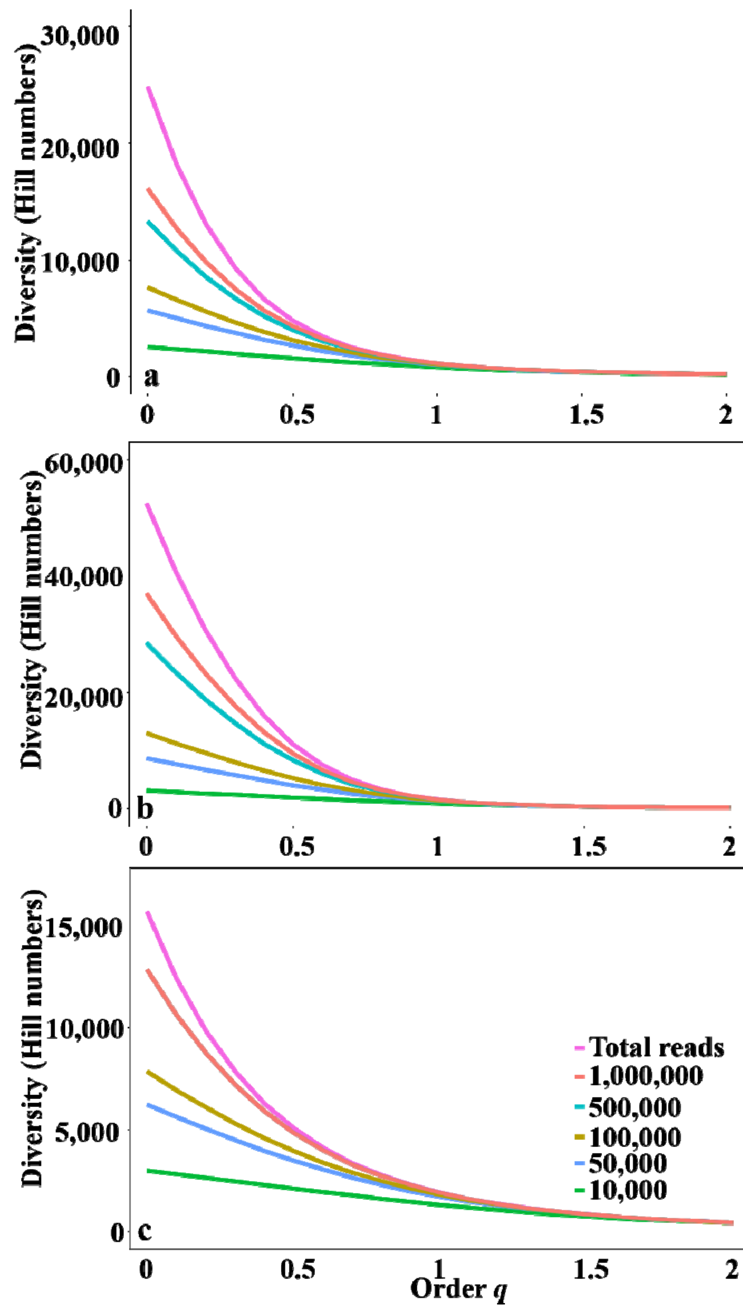

Fig. S4 Hill numbers profile for (a) UPARSE, (b) Deblur, and (c)DADA2. The x axis is the order q, $0 \leq q \leq 2$. The y axis is the Hill number.

Table S1 Soil properties of the four quadrats. AV, average values. SD, standard deviations.

Quadrat	Index	TN (mg/kg)	NH_3-N (mg/kg)	NO_3-N (mg/kg)	TOC (%)	Moisture content (%)	pH
Q1	AV	763.90	0.66	13.81	1.27	3.75	7.28
	SD	178.40	0.49	1.37	0.17	0.89	0.06
Q2	AV	669.23	0.79	11.02	1.18	3.74	7.25
	SD	129.42	0.46	2.21	0.18	0.83	0.11
Q3	AV	1400.64	1.39	9.63	2.54	4.11	7.27
	SD	242.12	1.91	1.65	0.29	1.00	0.11
Q4	AV	816.85	0.30	9.89	1.45	2.80	7.26
	SD	204.84	0.13	1.19	0.37	0.68	0.11

Table S2 Number of retained sequences after OTU or ASV generation.

Methods	Q1	Q2	Q3	Q4
UPARSE	10,993,500	9,543,092	7,681,474	8,864,097
Deblur	5,194,295	4,530,442	3,733,659	4,171,557
DADA2	10,815,483	9,402,775	7,697,276	9,097,673

Table S3 Relationships of sequencing depth to estimated richness by UPARSE.

Reads ($*10^4$)	Chao1 (%)	Chao2 (%)	ACE (%)	ICE (%)	Abundance_Jack1 (%)	Incidence_Jack1 (%)
5	33.18±1.30	33.38±1.32	26.76±1.25	28.30±1.41	28.80±1.38	29.66±1.45
10	41.78±1.27	41.64±1.23	34.75±1.39	33.70±1.44	37.57±1.51	37.93±1.55
20	51.29±1.46	50.95±1.47	43.76±1.47	43.59±1.57	47.23±1.60	47.29±1.70
30	57.32±1.68	56.78±1.51	49.62±1.58	49.39±1.50	53.34±1.69	53.18±1.60
50	64.88±1.80	64.43±1.69	57.36±1.62	57.23±1.59	61.25±1.70	61.04±1.66
80	72.13±1.89	71.37±1.81	64.91±1.73	64.72±1.73	68.80±1.77	68.40±1.76
100	75.44±1.98	74.64±1.78	68.55±1.85	68.34±1.78	72.36±1.86	71.91±1.78

Table S4 Relationships of sequencing depth to estimated richness by Deblur.

Reads ($*10^4$)	Chao1 (%)	Chao2 (%)	ACE (%)	ICE (%)	Abundance_Jack1 (%)	Incidence_Jack1 (%)
5	32.73±1.84	35.08±1.78	23.18±1.61	25.85±1.60	23.13±1.65	23.75±1.35
10	44.93±1.72	46.75±2.25	33.23±1.85	32.56±2.39	34.11±2.05	34.64±2.43
20	58.54±1.76	60.31±1.56	45.47±2.07	46.14±2.09	47.76±2.43	48.44±2.44
30	66.89±2.15	68.3±2.25	53.71±2.34	54.32±2.35	57.03±2.67	57.58±2.74
50	77.19±1.93	77.46±2.30	65.06±1.89	65.06±2.05	69.61±2.13	69.48±2.27
80	84.94±2.36	85.25±2.17	75.21±2.20	75.37±1.94	80.48±2.35	80.48±2.09
100	88.24±2.19	88.39±2.21	79.84±2.16	79.93±2.12	85.18±2.28	85.09±2.24

Table S5 Relationships of sequencing depth to estimated richness by DADA2.

Reads ($*10^4$)	Chao1 (%)	ACE (%)	Abundance_Jack1 (%)
5	42.27±1.74	36.79±1.39	42.09±1.66
10	52.95±1.99	47.05±1.79	53.22±2.04
20	63.97±1.78	58.11±1.86	64.69±1.97
30	69.74±1.90	63.99±1.95	70.51±2.06
50	77.11±1.91	71.66±2.06	77.90±2.07
80	83.49±1.95	78.31±1.95	84.16±1.90
100	87.15±2.23	81.82±2.06	87.59±2.06

Table S6 Species estimated by scaling laws: $S = 7.6*N^{0.35}$, where S is species richness and N is the number of individuals. 16S rDNA copies were counted by qPCR.

Quadrats	16S rDNA copies in 1g topsoil (0-20 cm)	Estimated number of prokaryotic cells	Estimated species in $1m^2$ topsoil (0-20 cm)
Q1	$1.0054*10^{10}$	$2.6342*10^{14}$	$8.4731*10^{5}$
Q2	$8.1851*10^{9}$	$2.1445*10^{14}$	$7.8846*10^{5}$
Q3	$2.0497*10^{10}$	$5.3702*10^{14}$	$1.0872*10^{6}$
Q4	$2.1062*10^{10}$	$5.5183*10^{14}$	$1.0976*10^{6}$
Average value	$1.4950*10^{10}$	$3.9168*10^{14}$	$9.7351*10^{5}$