

RESEARCH

Open Access

# Estimate of the sequenced proportion of the global prokaryotic genome



Zheng Zhang<sup>1,2\*†</sup> , Jianing Wang<sup>1†</sup>, Jinlan Wang<sup>3</sup>, Jingjing Wang<sup>1</sup> and Yuezhong Li<sup>1\*</sup>

## Abstract

**Background:** Sequencing prokaryotic genomes has revolutionized our understanding of the many roles played by microorganisms. However, the cell and taxon proportions of genome-sequenced bacteria or archaea on earth remain unknown. This study aimed to explore this basic question using large-scale alignment between the sequences released by the Earth Microbiome Project and 155,810 prokaryotic genomes from public databases.

**Results:** Our results showed that the median proportions of the genome-sequenced cells and taxa (at 100% identities in the 16S-V4 region) in different biomes reached 38.1% (16.4–86.3%) and 18.8% (9.1–52.6%), respectively. The sequenced proportions of the prokaryotic genomes in biomes were significantly negatively correlated with the alpha diversity indices, and the proportions sequenced in host-associated biomes were significantly higher than those in free-living biomes. Due to a set of cosmopolitan OTUs that are found in multiple samples and preferentially sequenced, only 2.1% of the global prokaryotic taxa are represented by sequenced genomes. Most of the biomes were occupied by a few predominant taxa with a high relative abundance and much higher genome-sequenced proportions than numerous rare taxa.

**Conclusions:** These results reveal the current situation of prokaryotic genome sequencing for earth biomes, provide a more reasonable and efficient exploration of prokaryotic genomes, and promote our understanding of microbial ecological functions.

**Keywords:** Microbiome, Genome sequencing, Prokaryotic biome, Earth microbiome project, Predominant taxa

## Background

Prokaryotes are generally assumed to be the oldest existing form of life on earth and the primary engines of global biogeochemical processes; they are found in almost all ecosystems [1, 2]. Genome sequencing provides a blueprint for the evolutionary and functional diversities of prokaryotes and improves our understanding of how they interact with one another, their hosts, and their surroundings [3–5]. However, what is the cells or taxa proportion of genome-sequenced bacteria or archaea on earth? This basic and seemingly simple question has never been answered.

Since the first bacterial genome was completely sequenced in 1995, more than 200,000 bacterial and archaeal complete or draft genomes have been uploaded to public databases as a result of the development of sequencing technology and the decrease in costs [6, 7]. Meanwhile, due to improvements in sequencing throughput and computational techniques, cultivation-independent recovery of genomes from metagenomes further promotes prokaryotic genome mining [8–10]. Interestingly, compared to the exponential accumulation of genomic data, the latest estimate of global prokaryotic operational taxonomic units (OTUs, 16S-V4 regions at 97% sequence identities) is only 0.8–1.6 million, far less than the trillions previously predicted [11, 12]. It is necessary to globally evaluate the proportion of sequenced prokaryotic genomes in environments.

\* Correspondence: [zhangzheng@sdu.edu.cn](mailto:zhangzheng@sdu.edu.cn); [lilab@sdu.edu.cn](mailto:lilab@sdu.edu.cn)

†Zheng Zhang and Jianing Wang contributed equally to this work.

<sup>1</sup>State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China

Full list of author information is available at the end of the article



© The Author(s). 2020 **Open Access** This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by/4.0/>. The Creative Commons Public Domain Dedication waiver (<http://creativecommons.org/publicdomain/zero/1.0/>) applies to the data made available in this article, unless otherwise stated in a credit line to the data.

The Earth Microbiome Project (EMP) was founded in 2010 to sample and explore the Earth's microbial communities at an unprecedented scale [13–15]. In this study, we conducted a large-scale sequence alignment between the data released by the EMP and the sequenced bacterial or archaeal genomes in the public database. From these data, we evaluated the present situation of prokaryotic genome sequencing in the earth biomes for the first time.

## Results

### High genome-sequenced proportions in different prokaryotic biomes

A representative subset, containing 10,000 samples to represent different environment types, was selected from 27,751 samples of 97 independent studies released by the EMP [13].  $B_{\text{cell}}$  and  $B_{\text{OTU}}$ , which represent the genome-sequenced proportions of cells and taxa (at 100%, > 98.6%, or > 97% identities in the 16S-V4 region) in a specific prokaryotic biome, respectively, were evaluated based on the alignment between the 16S rRNA gene sequences of the EMP and the nearly 155,810 RefSeq genome sequences. The results showed that the median  $B_{\text{cell (100%)}}$  in the 10,000 samples was 38.1%, and the upper and lower quartiles were 16.4% and 86.3%, respectively (Fig. 1a and Supplementary dataset 1). This finding indicates that the genome information of at least 38% of cells has been reported in more than half of the prokaryotic biomes. The median  $B_{\text{OTU (100%)}}$  reached 18.8% (9.1–52.6%) (Fig. 1b). Generally, closely related strains with high similarities of 16S rRNA gene sequences (97% or 98.6%) also share high genome similarity [16–18]. The median  $B_{\text{cell (98.6%)}}$  was 50.1% (28.3–90.6%) whereas the median  $B_{\text{cell (97%)}}$  reached 60.4% (40.0–93.0%) across the 10,000 samples. Similarly, the median  $B_{\text{OTU (98.6%)}}$  was 28.4% (16.4–64.3%), and the median  $B_{\text{OTU (97%)}}$  increased to 37.6% (24.8–71.4%) (Supplementary Figs. S1, S2).

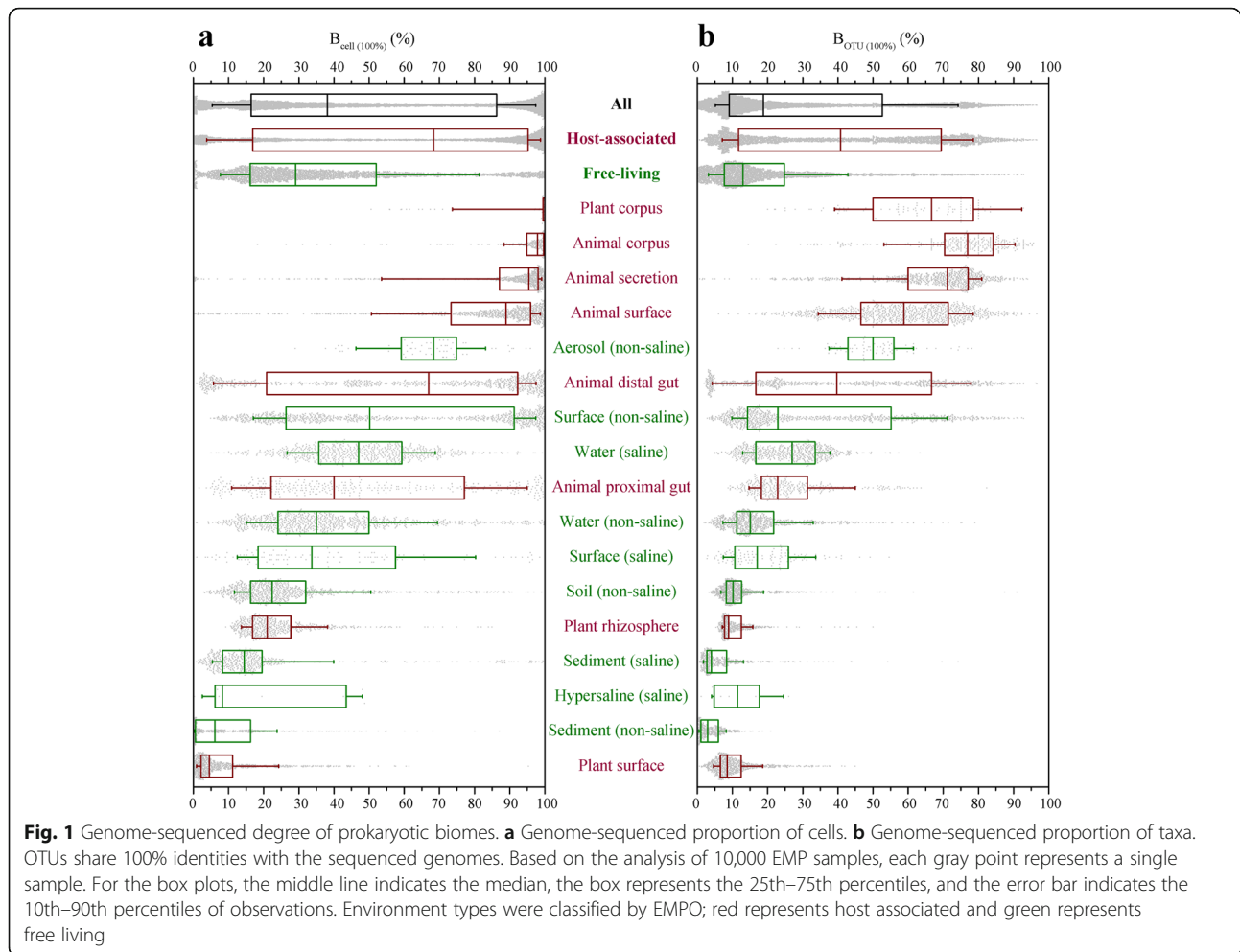
The genome-sequenced proportion in the prokaryotic biome was closely related to habitat (Fig. 1 and Supplementary dataset 2). Microbial environments are divided into different environment types by the EMP. The EMP ontology (EMPO level 1) classifies microbial environments as free living and host associated, with further subdivision into 17 environment types (EMPO level 3) [13]. We found that the genome-sequenced proportions in host-associated biomes were significantly higher than those in free-living biomes. For the host-associated prokaryotic biomes (5161 samples), the median  $B_{\text{cell (100%)}}$  was as high as 68.3% (16.9–95.2%), and the median  $B_{\text{OTU (100%)}}$  was 40.7% (11.8–69.4%). However, for the free-living prokaryotic biomes (4839 samples), the median  $B_{\text{cell (100%)}}$  was only 29.1% (16.2–52.0%), and the median  $B_{\text{OTU (100%)}}$  was 13.0% (7.7–24.8%). In detail, the median

$B_{\text{cell (100%)}}$  in plant corpus, animal corpus, and animal secretions exceeded 95%, and the median  $B_{\text{OTU (100%)}}$  exceeded 66.7%. Comparatively, the median  $B_{\text{cell (100%)}}$  values for plant surface, sediment (non-saline), and hypersaline samples were all less than 10%, and the median  $B_{\text{OTU (100%)}}$  values for sediment (non-saline) and sediment (saline) samples were less than 5% (Fig. 1). For closely related strains,  $B_{\text{cell}}$  and  $B_{\text{OTU}}$  also showed similar variabilities among different habitats (Supplementary Figs. S1, S2). Despite significant differences, the genome-sequenced proportions were high in most of the prokaryotic biomes.

Furthermore, we found that the genome-sequenced proportion in the prokaryotic biome was significantly negatively correlated with its alpha diversity indices (Supplementary Fig. S3). For both cells and taxa, the prokaryotic biomes with low alpha diversity indices (observed OTUs, Shannon index, Chao1 index, and Faith's PD value) tended to have a higher degree of genome sequencing. For example, the Pearson correlation coefficients of  $B_{\text{cell (100%)}}$  and  $B_{\text{OTU (100%)}}$  with Shannon indices were  $-0.62$  ( $p < 0.01$ ) and  $-0.67$  ( $p < 0.01$ ), respectively.

### Low genome-sequenced proportions of global prokaryotic taxa

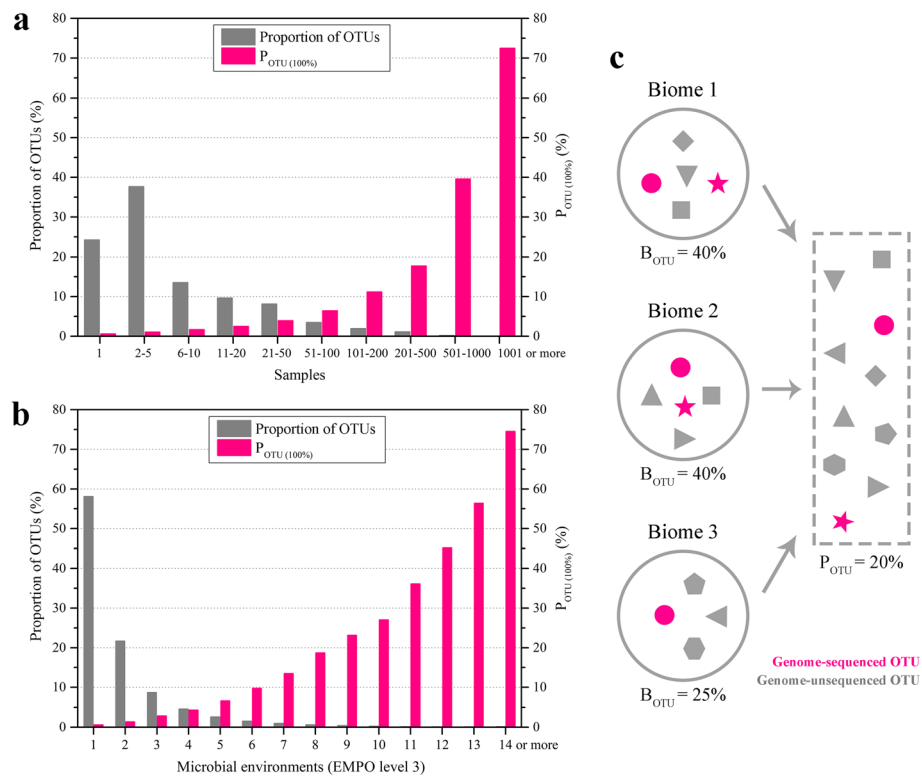
A total of 262,011 OTUs were obtained from 10,000 EMP samples through a meta-analysis. We defined the genome-sequenced proportion of all taxa (at 100%, > 98.6%, or > 97% identities in the 16S-V4 region) as  $P_{\text{OTU}}$  and found that the  $P_{\text{OTU (100%)}}$  of the 10,000 samples was only 2.1% (Supplementary dataset 3). The  $P_{\text{OTU (98.6%)}}$  and  $P_{\text{OTU (97%)}}$  values were 6.8% and 12.2%, respectively, and both were also much lower than the corresponding  $B_{\text{cell}}$  and  $B_{\text{OTU}}$  medians. Furthermore, we found that 75.8% of OTUs were present in two or more biome samples. The  $P_{\text{OTU (100%)}}$  value was 0.6% for the OTUs that appeared in only a single sample (401 of 63,459 OTUs), 1.2% for those in 2 to 10 samples (1641 of 134,119 OTUs), 5.4% for those in more than 10 samples (3478 of 64,433 OTUs), 16.2% for those in more than 100 samples (1431 of 8810 OTUs), and 72.5% for those in more than 1000 samples (108 of 149 OTUs) (Fig. 2a). Notably, many prokaryotic taxa could exist in diverse environment types; approximately 21.7% of prokaryotic taxa could exist in two types of environments, and 20.2% of OTUs could exist in three or more types of environments. We found that the taxon genome-sequenced proportion also increased with its distribution extent in different environment types. The  $P_{\text{OTU (100%)}}$  was only 0.6% for prokaryotic OTUs that existed in only one type of environment (932 of 152,229 OTUs), 14.5% for OTUs in five or more types of environments (2645 of 18,230 OTUs), 43.6% for 10 or more types of environments



(904 of 2074 OTUs), and 74.6% for 14 or more types of environments (287 of 385 OTUs) (Fig. 2b). A higher genome-sequenced proportion of prokaryotic cosmopolitan OTUs led to a lower  $P_{OTU}$  than the corresponding  $B_{OTU}$  (Fig. 2c).

Because an OTU was likely to appear in multiple samples, we evaluated the effects of sample quantity on  $P_{OTU}$  by random sampling. Our results demonstrated that the  $P_{OTU(100\%)}$  displayed an exponential decay trend ( $R^2 = 0.992$ ) and eventually stabilized at  $2.13\% \pm 0.03\%$  as the number of samples increased (Fig. 3a). Similarly, the  $P_{OTU(98.6\%)}$  and  $P_{OTU(97\%)}$  values also decreased with increasing sample size and stabilized at approximately 6.8% and 12.2%, respectively (Supplementary Fig. S4). The estimated  $P_{OTU}$  values based on 10,000 EMP samples were close to the genome-sequenced proportions in all global prokaryotic taxa. We evaluated the changes in  $P_{OTU}$  as the number of sequenced genomes increased from 2010 to 2019. The results showed that the  $P_{OTU(100\%)}$  increased

exponentially ( $R^2 = 0.998$ ) by sixfold over the decade. However, it was estimated that it would take at least 25 years for the  $P_{OTU(100\%)}$  to reach 95%. With the increase in sequenced genomes, the  $P_{OTU(100\%)}$  value showed an allometric increase ( $R^2 = 0.989$ ), and we determined that the 95%  $P_{OTU(100\%)}$  value required more than  $10^9$  sequenced genomes (Supplementary Fig. S5). In addition, the  $P_{OTU}$  also differed significantly between environments. The  $P_{OTU(100\%)}$  value based on the total host-associated samples was 4.6% whereas the  $P_{OTU(100\%)}$  value for all the free-living samples was only 2.1%. The  $P_{OTU(100\%)}$  values for the animal corpus and plant corpus environments were 28.3% and 23.7%, respectively, whereas the  $P_{OTU(100\%)}$  values for sediment (non-saline), soil (non-saline), and water (non-saline) environments were only 2.3%, 2.9%, and 2.9%, respectively.  $P_{OTU(98.6\%)}$  and  $P_{OTU(97\%)}$  also showed similar patterns (Fig. 3b). Thus, despite the rapid accumulation of prokaryotic genomic information, the genome-sequenced proportion of the global prokaryotic taxa was still fairly low.



**Fig. 2** High genome-sequenced proportion of prokaryotic cosmopolitan taxa. **a** OTUs that can exist in one or more samples. **b** OTUs that can exist in one or more environment types. The gray column represents the proportion of OTUs that can exist in one or more samples (environments), and the red column represents the genome-sequenced proportion of OTUs. **c** Lower  $P_{OTU}$  than  $B_{OTU}$  is caused by a high genome-sequenced proportion of cosmopolitan taxa

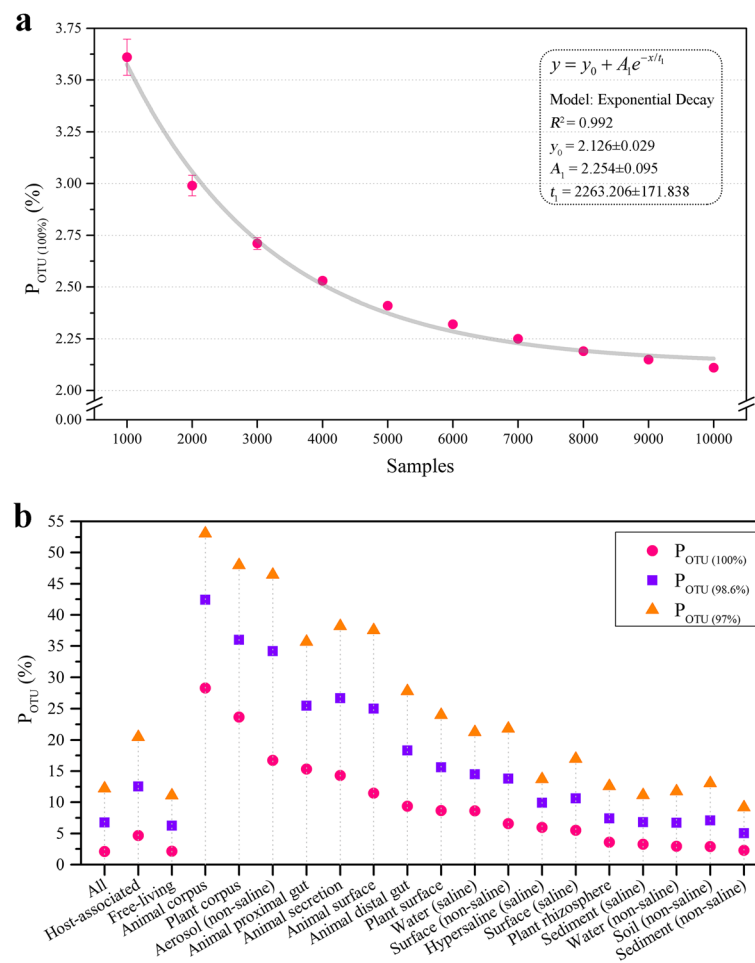
### The majority of the biomes were occupied by a few predominant taxa with high relative abundances

Our results showed that the top 1% of the prokaryotic taxa (sorted by their percentage of 16S rRNA sequences) accounted for 72.9% of the global prokaryotic biomes (Fig. 4a and Supplementary Fig. S6). These top 1% of taxa always had a high abundance in different environment types (Fig. 4b), which was similar to a recent report on global soil dominant bacteria [19]. By contrast, the rare taxa with low abundance (the total number of sequences < 10) accounted for 59.8% of the total prokaryotic taxa but only 1.2% of the global prokaryotic cells (Supplementary Fig. S7). We found that the number of samples affected the observed proportion of rare taxa to global taxa; as the number of samples increased, the ratio value increased gradually (Supplementary Fig. S8). Notably, the genome-sequenced proportion of the top 1% of prokaryotic taxa reached 38.0% whereas that of the 59.8% of prokaryotic taxa with a low abundance was only 0.6% (Fig. 4c and Supplementary Fig. S6). The genome-sequenced proportions of the top 1% of prokaryotic taxa from different environment types exceeded 12% (Fig. 4b). We further selected 1325 highly

abundant and widely distributed OTUs on the following conditions: existing in at least 9 environment types and at least 100 samples and had an abundance reaching the top 1% in at least 1 type of environment (Supplementary dataset 3). These predominant taxa accounted for only 0.5% of the total OTUs but contributed to 50.3% of the global prokaryotic biomes. The genome-sequenced proportion was fairly high in these dominant taxa, and the  $P_{OTU(100\%)}$ ,  $P_{OTU(98.6\%)}$ , and  $P_{OTU(97\%)}$  values were 48.2%, 61.7%, and 71.3%, respectively (Supplementary Fig. S9). The majority of biomes were occupied by a few predominant taxa with high genome-sequenced proportions.

### Culturability altered genome-sequenced preferences among prokaryotes but not environments

We estimated the  $P_{OTU}$  values of prokaryotes at different taxonomic levels (Supplementary dataset 4), which showed that the  $P_{OTU}$  values were obviously different among different taxa, and the  $P_{OTU}$  value of the same taxon also differed significantly among different environment types (Supplementary Figs. S10, S11, S12, S13, S14 and S15). For example, of the 11 phyla with OTU



**Fig. 3** Genome-sequenced proportion of prokaryotic taxa from global or different environment types. **a** As the number of samples increases, the  $P_{OTU (100\%)}$  shows an exponential declining trend and finally stabilizes at 2.1%. A random selection of 1000, 2000..., 9000 samples was performed 10 times for each group to calculate the mean value and standard deviation. **b** Significant difference of  $P_{OTU}$  among environment types. The red point is  $P_{OTU (100\%)}$ , the blue point is  $P_{OTU (98.6\%)}$ , and the orange point is  $P_{OTU (97\%)}$

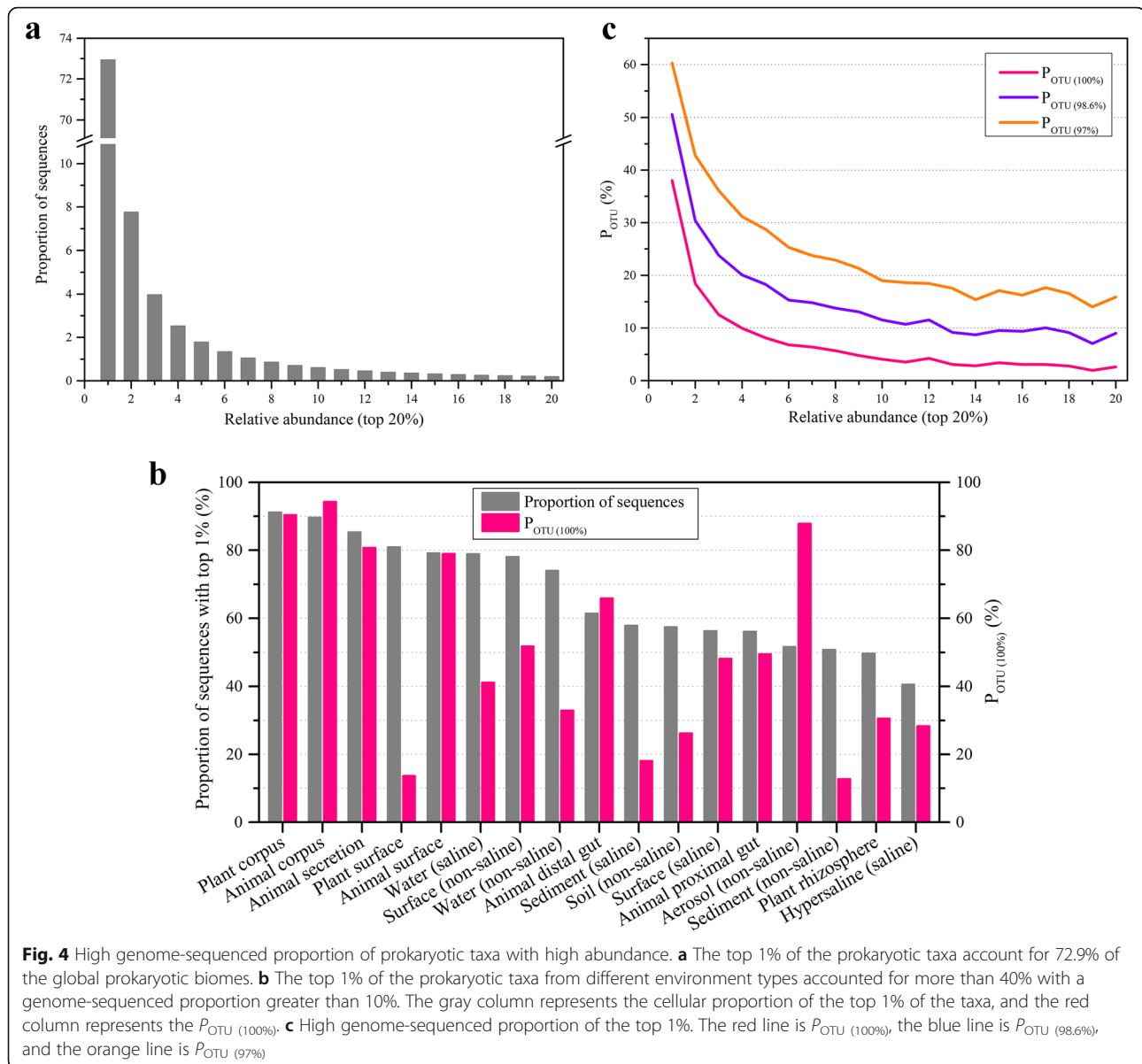
numbers greater than 1%, the highest  $P_{OTU (100\%)}$  value was 5.7% for *Actinobacteria*, and the lowest  $P_{OTU (100\%)}$  value was 0.04% for *Parcubacteria*; the difference between them spanned more than 100-fold (Supplementary Fig. S10).

Due to improvements in sequencing throughput and computational techniques, cultivation-independent recovery of genomes from metagenomic data has rapidly developed. In total, 7903 bacterial and archaeal metagenome-assembled genomes (MAGs) were recovered from massive metagenomic data, which were considered from uncultivated strains [8]. We assessed the effect of strain culturability on the current genomic sequencing preferences using these MAGs and 155,810 cultured genomes (Supplementary dataset 5). The results showed that the genome-sequenced proportion of prokaryotes increased by 0.1% after combining these MAGs. According to the environment types, the  $P_{OTU}$

(100%) based on MAGs was highly positively correlated with that based on RefSeq ( $r = 0.91$ ,  $p < 0.01$ ) (Supplementary Fig. S16). The result showed that, similar to the RefSeq genomes, the MAGs also showed environmental differences, and the culturability of strains was not the main factor leading to these differences. For the 11 phyla with an OTU number proportion greater than 1%, there was no significant correlation between the  $P_{OTU (100\%)}$  based on the MAGs and the RefSeq ( $p > 0.05$ ) (Supplementary Fig. S16). This indicated that although the recovered MAGs had a distinct difference in prokaryotic taxa, its species preference was significantly different from the RefSeq genomes.

## Discussion

The genome is the basic resource for understanding the physiology, ecology, and evolution of prokaryotes. More than 200,000 bacterial and archaeal genomes are now



available from over two decades of development [3, 6]. These genomes provide important insights into the role of microorganisms in industrial processes, the pathogenic mechanisms of pathogenic microorganisms, etc. In this study, we assessed the genome-sequenced proportion of global prokaryotes. We found that the median proportions of the genome-sequenced prokaryotic cells and taxa (at 100% identities in the 16S-V4 region) in global biomes were 38.1% (16.4–86.3%) and 18.8% (9.1–52.6%), respectively. The  $B_{cell(97\%)}$  of 61.9% of the samples reached 50%, and the  $B_{OTU(97\%)}$  of 38.4% of the samples reached 50% after combining closely related strains. In addition, the median  $B_{cell(97\%)}$  and  $B_{OTU(100\%)}$  values in host-associated biomes were 85.6% (43.2–98.0%) and 62.8% (9.8–82.3%), respectively, which

were significantly higher than those in free-living biomes. Thus, the genetic information of a specific prokaryotic biome may have been reported to a considerable degree.

However, compared to prokaryotic biomes, the genome-sequenced proportion of global prokaryotic OTUs was fairly low. Our results suggest that only 2.1% of the global prokaryotic taxa (at 100% identities in the 16S-V4 region) have been sequenced. More than 75% of prokaryotic OTUs could exist in multiple biomes; the more types of environments in which prokaryotic OTUs can survive, the higher the genome-sequenced proportion could be. Prokaryotic biomes are usually composed of a few predominant taxa with a high abundance and many rare taxa with a low abundance [20, 21]. We found

that 0.5% of predominant OTUs occupied 50.3% of prokaryotic cell abundance with a high genome-sequenced proportion (48.2%); however, the 60% of rare OTUs only accounted for 1.2% of the global prokaryotic cells with a low genome-sequenced proportion (0.6%). A large number of rare taxa are considered to be critical components of the earth's ecosystem and contain a large functional genes pool [21, 22]. Therefore, from this perspective, our current understanding of global prokaryotic genomic information remains very limited due to the large number of genome-unsequenced rare taxa, and the exploration of this huge genetic resource is just beginning.

Predominant taxa are considered the priority for isolated culture and genome sequencing [19]. We identified 1325 predominant OTUs with a wide distribution, high abundance, and adaptability to a variety of environmental types, more than half of which had not been genome-sequenced. In particular, some predominant taxa acquired less attention in specific environmental types. For example, the top 1% taxa of abundance in plant surfaces and animal surfaces accounted for 81.0% and 79.2% of the global prokaryotic biomes whereas the genome-sequenced proportions of the taxa were only 13.7% and 79.0%, respectively. The  $P_{\text{OTU}}(100\%)$  of plant surfaces (leaf or kelp surface biofilms) was ranked 8th, but its median  $B_{\text{cell}}$  was last given the lack of understanding of predominant taxa.

Currently, most of the prokaryotic sequenced genomes (RefSeq genomes) are from pure cultures, while MAGs are not limited by culturability [8, 9, 23]. We found similar genome-sequenced differences among different environment types between RefSeq genomes and MAGs, which indicated that the current imbalance of prokaryotic genome sequencing in different environments was more likely due to differences in researchers' attention rather than prokaryotic culturability. Although the significant genome-sequenced differences among different taxa between RefSeq genomes and MAGs suggested that culturability caused genomic sequencing preferences had no effect on MAGs, MAGs had also owned its own taxa sequenced preferences.

The paradigm that only 1% of prokaryotes are culturable has a profound impact on microbial ecology but has recently been debated [24–26]. Since the RefSeq genomes are mainly from culturable taxa, and a significant proportion of culturable taxa have not been sequenced, we estimate that the culturable rate of global prokaryotic taxa (> 97% identities) would be higher than the genome-sequenced proportion of 12.2%. Similar to the higher genome-sequenced proportion of the high abundance predominant taxa, predominant taxa should also have a much higher culturability rate than rare taxa; thus, the culturability rate of prokaryotic cells will be much higher than that of taxa. Consequently, our data

indicated that the paradigm that only 1% of prokaryotes are culturable is out of date, both for cells and taxa.

## Conclusions

This study performed an in-depth analysis of the prokaryotic genome-sequenced proportion in the EMP and comprehensively showed the global-scale genome-sequenced degree for various environment types and different species. Most of the biomes were occupied by a few widespread predominant taxa. Given the high genome-sequenced proportion of predominant taxa, the genetic information of most prokaryotic biomes has been revealed to a high degree. However, due to the large number of rare taxa with unknown genomes, our current understanding of the global prokaryotic genome information remains limited. These results will be helpful for more reasonable and efficient explorations of prokaryotic genomes and will accelerate the comprehensive understanding of microbial ecological functions in different environments.

## Methods

### Data collection from EMP and RefSeq

The Earth Microbiome Project (EMP) was founded in 2010 to sample the Earth's microbial communities at an unprecedented scale to advance our understanding of the organizing biogeographic principles that govern microbial community structure on Earth [13–15]. A total of 262,011 OTUs and their abundance and nucleic acid sequence information were collected from the website (<ftp://ftp.microbio.me/emp/release1>), which were obtained and shared by the EMP from 10,000 samples using the Deblur software [27]. Chimera filtering relied on the EMP project. The NCBI's reference sequence (RefSeq) database is a curated non-redundant collection of sequences representing whole or frame genomes [28]. We obtained all of the 155,810 bacterial or archaeal genomes collected by the database before July 2019. In addition, 7903 (1539 contained the 16S rRNA gene) metagenome-assembled genomes (MAGs) [8] recovered from > 1500 public metagenomes using MetaBAT [29] were also collected for representative uncultivated bacteria and archaea.

### Sequence alignment and analysis

Alignment between the EMP OTUs and 155,810 or 7903 genomes was performed using BLASTn ( $E$  value <  $1e-5$ ) [30]. To assess the adequacy of the OTUs, we analyzed all the samples by increasing the number of samples from 1000 to 10,000 randomly. The genome-sequenced proportions of cells and taxa (at 100%, > 98.6% or, > 97% identities in the 16S-V4 region) in a specific prokaryotic biome were defined as  $B_{\text{cell}}$  and  $B_{\text{OTU}}$ , respectively. The genome-sequenced proportion of taxa (at 100%, > 98.6%, or > 97% identities in the 16S-V4 region) from subgroup

or global biomes was defined as the  $P_{OTU}$ . The 100% identity represents the most rigorous and accurate match, while 98.6% and 97% identities are the new and traditional criteria for species definitions, respectively [16–18]. Briefly,  $B_{cell}$  represents the ratio of the genome-sequenced sequences in a single sample,  $B_{OTU}$  represents the ratio of the genome-sequenced OTUs in a single sample and  $P_{OTU}$  represents the ratio of the genome-sequenced OTUs in multiple samples.

### Taxonomic analysis of EMP OTU

The taxonomy of each OTU was analyzed by the Ribosomal Database Project (RDP) Classifier [31] at a 70% confidence threshold. The EMP ontology (EMPO) classified 17 microbial environments (level 3) as free living or host associated (level 1) and saline or non-saline (if free living) or animal or plant (if host associated) (level 2) [13]. Based on the taxonomic results and the EMPO (level 3) for each OTU, we calculated the composition and relative abundance of different levels of taxonomy (phylum, class, order, family, and genus) in different environments.

### Supplementary information

Supplementary information accompanies this paper at <https://doi.org/10.1186/s40168-020-00903-z>.

**Additional file 1 : Supplementary Fig. S1.** Genome-sequenced degree of prokaryotic biomes. **a**, Genome-sequenced proportion of cells. **b**, Genome-sequenced proportion of taxa. OTUs share at least 98.6% identities with the sequenced genomes. Based on the analysis of 10,000 EMP samples, each grey point represents a single sample. For the box plots, the middle line indicates the median, the box represents the 25th–75th percentiles, and the error bars represent the 10th–90th percentiles of observations. Environment types were classified by EMPO; red represents host-associated and green represents free-living.

**Additional file 2: Supplementary Fig. S2.** Genome-sequenced degree of prokaryotic biomes. **a**, Genome-sequenced proportion of cells. **b**, Genome-sequenced proportion of taxa. OTUs share at least 97% identities with the sequenced genomes. Based on the analysis of 10,000 EMP samples, each grey point represents a single sample. For the box plots, the middle line indicates the median, the box represents the 25th–75th percentiles, and the error bars indicate the 10th–90th percentiles of observations. Environment types were classified by EMPO; red represents host-associated and green represents free-living.

**Additional file 3: Supplementary Fig. S3.** Genome-sequenced proportions of prokaryotic biomes are significantly negatively correlated with the biomes' alpha diversity indices. Alpha diversity indices include observed OTUs, Shannon index, Chao1 index, and Faith's PD value. Each point represents one single sample from a total of 10,000 samples. Brown represents  $B_{cell}$  (100%) and purple represents  $B_{OTU}$  (100%).

**Additional file 4: Supplementary Fig. S4.** Genome-sequenced proportion of global prokaryotic taxa. As the number of samples increases, the  $P_{OTU}$  (98.6%) and  $P_{OTU}$  (97%) values show an exponential declining trend and finally stabilize at 6.8% and 12.2%, respectively. A random selection of 1000, 2000..., 9000 samples was performed 10 times for each group to calculate the mean value and standard deviation. Blue represents  $P_{OTU}$  (98.6%) and orange represents  $P_{OTU}$  (97%).

**Additional file 5: Supplementary Fig. S5.** The trends of  $P_{OTU}$  over time and genome number. **a**, The  $P_{OTU}$  has grown exponentially over time. **b**, The  $P_{OTU}$  shows an allometric rising trend as the number of

sequenced genomes increases. Red represents  $P_{OTU}$  (100%), blue represents  $P_{OTU}$  (98.6%) and orange represents  $P_{OTU}$  (97%).

**Additional file 6: Supplementary Fig. S6.** The top 1% of the prokaryotic taxa accounts for 72.9% of the global prokaryotic biomes with high genome-sequenced proportion. The red line is  $P_{OTU}$  (100%), the blue line is  $P_{OTU}$  (98.6%), and the orange line is  $P_{OTU}$  (97%).

**Additional file 7: Supplementary Fig. S7.** The rare taxa with low abundance (total number of sequences < 10) account for 59.8% of the total prokaryotic taxa but only 1.2% of the global prokaryotic cells with a 0.6% genome-sequenced proportion.

**Additional file 8: Supplementary Fig. S8.** The proportion of rare taxa to the global taxa increases gradually and stabilizes at approximately 60% as the number of samples increases.

**Additional file 9: Supplementary Fig. S9.** Predominant taxa have a high abundance and wide distribution. **a**, A total of 1,325 OTUs were selected according to the following conditions: existing in at least 9 environments and at least 100 samples and an abundance reaching the top 1% in at least 1 environmental type. **b**, OTU number proportion and  $P_{OTU}$  (100%) (in brackets) of the main phylum in the predominant taxa. **c**, OTU number proportion and  $P_{OTU}$  (100%) (in brackets) of the main class in the predominant taxa. **d**, OTU number proportion and  $P_{OTU}$  (100%) (in brackets) of the main order in the predominant taxa.

**Additional file 10: Supplementary Fig. S10.** Obviously genome-sequenced preferences of prokaryotes among taxa. **a**, The OTU number proportion of the 11 main phyla. **b**, The  $P_{OTU}$  (100%) of the 11 main phyla.

**Additional file 11: Supplementary Fig. S11.** Heatmap of  $P_{OTU}$  (100%) between the 11 main phyla and the 17 environment types.

**Additional file 12: Supplementary Fig. S12.** Genome-sequenced proportion of 46 main classes of prokaryotic predominant taxa. The parenthesis shows the OTU number proportion of each class in all prokaryotic OTUs.

**Additional file 13: Supplementary Fig. S13.** Genome-sequenced proportion of the 49 main orders of prokaryotic predominant taxa. The parenthesis shows the OTU number proportion of each order in all prokaryotic OTUs.

**Additional file 14: Supplementary Fig. S14.** Genome-sequenced proportion of the 58 main families of prokaryotic predominant taxa. The parenthesis shows the OTU number proportion of each family in all prokaryotic OTUs.

**Additional file 15: Supplementary Fig. S15.** Genome-sequenced proportion of 55 main genera of prokaryotic predominant taxa. The parenthesis shows the OTU number proportion of each genus in all prokaryotic OTUs.

**Additional file 16: Supplementary Fig. S16.** Culturability alters genome-sequenced preferences among prokaryotes but not environments. **a**, For the 17 environment types, the  $P_{OTU}$  (100%) based on MAGs is highly positively correlated with that based on RefSeq ( $r = 0.91$ ,  $p < 0.01$ ). **b**, For the 11 phyla with an OTU number proportion greater than 1%, the  $P_{OTU}$  (100%) based on the MAGs has no significant correlation with that based on the RefSeq ( $p > 0.05$ ).

**Additional file 17: Supplementary dataset 1.**  $B_{cell}$ ,  $B_{OTU}$ , alpha diversity index and environment type of 10,000 samples.

**Additional file 18: Supplementary dataset 2.**  $B_{cell}$ ,  $B_{OTU}$  and  $P_{OTU}$  of different environment types.

**Additional file 19: Supplementary dataset 3.** Summary of 262,011 OTUs and predominant taxa.

**Additional file 20: Supplementary dataset 4.** The  $P_{OTU}$  of different taxonomic levels (domain, phylum, class, order, family and genus) in different environment types.

**Additional file 21: Supplementary dataset 5.** The  $P_{OTU}$  of prokaryote-based MAGs.

### Acknowledgements

We thank Peng Zhang, Li Zhuo, and Ya Liu for their helpful discussion.



**Authors' contributions**

ZZ, JNW, and YL designed the project. ZZ, JNW, JLW, and JJW performed the experiments. ZZ, JNW, JLW, JJW, and YL analyzed the results. ZZ, JNW, and YL wrote the manuscript. The authors read and approved the final manuscript.

**Funding**

This work was financially supported by the National Natural Science Foundation of China (31670076), the Special Investigation on Scientific and Technological Basic Resources (2017FY100300), the National Key Research and Development Program (2018YFA0900400 and 2018YFA0901704), the Key Research and Development Program of Shandong Province (2018GSF121015) (YL), the Natural Science Foundation of Jiangsu Province (BK20190199), the China Postdoctoral Science Foundation (2018M642649 and 2019T120586), the Special Funding for Postdoctoral Innovation Project of Shandong Province (201902014), and the Qingdao Postdoctoral Application Research Project (ZZ).

**Availability of data and materials**

The main data supporting the findings of this study are available within the article and in its Supplementary Information. All other data supporting the findings of this study are available from the corresponding authors upon reasonable request.

**Ethics approval and consent to participate**

Not applicable

**Consent for publication**

Not applicable

**Competing interests**

The authors declare no competing interests.

**Author details**

<sup>1</sup>State Key Laboratory of Microbial Technology, Institute of Microbial Technology, Shandong University, Qingdao 266237, China. <sup>2</sup>Suzhou Research Institute, Shandong University, Suzhou 215123, China. <sup>3</sup>Physical Examination Office of Shandong Province, Health Commission of Shandong Province, Jinan 250014, China.

Received: 5 February 2020 Accepted: 28 July 2020

Published online: 16 September 2020

**References**

- Fuhrman JA, Cram JA, Needham DM. Marine microbial community dynamics and their ecological interpretation. *Nat Rev Microbiol*. 2015;13(3):133–46.
- Fierer N. Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol*. 2017;15(10):579–90.
- Loman NJ, Pallen MJ. Twenty years of bacterial genome sequencing. *Nat Rev Microbiol*. 2015;13(12):787–94.
- Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN, et al. A phylogeny-driven genomic encyclopaedia of bacteria and archaea. *Nature*. 2009;462(7276):1056–60.
- Fraser CM, Eisen JA, Salzberg SL. Microbial genome sequencing. *Nature*. 2000;406(6797):799–803.
- Land M, Hauser L, Jun SR, Nookaew I, Leuze MR, Ahn TH, et al. Insights from 20 years of bacterial genome sequencing. *Funct Integr Genomics*. 2015;15(2):141–61.
- Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, et al. Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd. *Science*. 1995;269(5223):496–512.
- Parks DH, Rinke C, Chuvochina M, Chaumeil PA, Woodcroft BJ, Evans PN, et al. Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat Microbiol*. 2017;2(11):1533–42.
- Pasolli E, Asnicar F, Manara S, Zolfo M, Karcher N, Armanini F, et al. Extensive unexplored human microbiome diversity revealed by over 150,000 genomes from metagenomes spanning age, geography, and lifestyle. *Cell*. 2019;176(3):649–62 e20.
- Stewart RD, Auffret MD, Warr A, Walker AW, Roehe R, Watson M. Compendium of 4,941 rumen metagenome-assembled genomes for rumen microbiome biology and enzyme discovery. *Nat Biotechnol*. 2019;37(8):953–61.
- Louca S, Mazel F, Doebeli M, Parfrey LW. A census-based estimate of Earth's bacterial and archaeal diversity. *PLoS Biol*. 2019;17(2):e3000106.
- Locey KJ, Lennon JT. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci U S A*. 2016;113(21):5970–5.
- Thompson LR, Sanders JG, McDonald D, Amir A, Ladau J, Locey KJ, et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature*. 2017;551(7681):457–63.
- Gilbert JA, Meyer F, Antonopoulos D, Balaji P, Brown CT, Brown CT, et al. Meeting report: the terabase metagenomics workshop and the vision of an earth microbiome project. *Stand Genomic Sci*. 2010;3(3):243–8.
- Gilbert JA, Jansson JK, Knight R. The earth microbiome project: successes and aspirations. *BMC Biol*. 2014;12:69.
- Stackebrandt E, Goebel BM. Taxonomic note: a place for DNA-DNA reassociation and 16S rRNA sequence analysis in the present species definition in bacteriology. *Int J Syst Evol Microbiol*. 1994;44(4):846–9.
- Rossello-Mora R, Amann R. Past and future species definitions for bacteria and archaea. *Syst Appl Microbiol*. 2015;38(4):209–16.
- Chun J, Oren A, Ventosa A, Christensen H, Arahal DR, da Costa MS, et al. Proposed minimal standards for the use of genome data for the taxonomy of prokaryotes. *Int J Syst Evol Microbiol*. 2018;68(1):461–6.
- Delgado-Baquerizo M, Oliverio AM, Brewer TE, Benavent-Gonzalez A, Eldridge DJ, Bardgett RD, et al. A global atlas of the dominant bacteria found in soil. *Science*. 2018;359(6373):320–5.
- Nemergut DR, Costello EK, Hamady M, Lozupone C, Jiang L, Schmidt SK, et al. Global patterns in the biogeography of bacterial taxa. *Environ Microbiol*. 2011;13(1):135–44.
- Jousset A, Bienhold C, Chatzinotas A, Gallien L, Gobet A, Kurm V, et al. Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J*. 2017;11(4):853–62.
- Lynch MD, Neufeld JD. Ecology and exploration of the rare biosphere. *Nat Rev Microbiol*. 2015;13(4):217–29.
- Solden L, Lloyd K, Wrighton K. The bright side of microbial dark matter: lessons learned from the uncultivated majority. *Curr Opin Microbiol*. 2016;31:217–26.
- Martiny AC. High proportions of bacteria are culturable across major biomes. *ISME J*. 2019;13(8):2125–8.
- Steen AD, Crits-Christoph A, Carini P, DeAngelis KM, Fierer N, Lloyd KG, et al. High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J*. 2019;13:3126.
- Puspita ID, Kamagata Y, Tanaka M, Asano K, Nakatsu CH. Are uncultivated bacteria really uncultivable? *Microbes Environ*. 2012;27(4):356–66.
- Amir A, McDonald D, Navas-Molina JA, Kopylova E, Morton JT, Zech Xu Z, et al. Deblur rapidly resolves single-nucleotide community sequence patterns. *mSystems*. 2017;2(2).
- Pruitt KD, Tatusova T, Maglott DR. NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*. 2007;35(Database issue):D61–5.
- Kang DD, Froula J, Egan R, Wang Z. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*. 2015;3:e1165.
- Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.
- Cole JR, Wang Q, Fish JA, Chai B, McGarrell DM, Sun Y, et al. Ribosomal database project: data and tools for high throughput rRNA analysis. *Nucleic Acids Res*. 2014;42(Database issue):D633–42.

**Publisher's Note**

Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.