environmental microbiology

Environmental Microbiology (2021) **00**(00), 00–00 doi:10.1111/1462-2920.15641

A few dominant bacteria and their genomic basis in mediating distinct ecosystem functions

Minglei Ren [1] **and Jianjun Wang** [1,2*]

[1]*State Key Laboratory of Lake Science and Environment, Nanjing Institute of Geography and Limnology, Chinese Academy of Sciences, Nanjing, 210008, China.*

[2]*University of Chinese Academy of Sciences, Beijing, 100049, China.*

Summary

Species attributes such as abundance and traits are important determinant components for ecosystem functions (EFs), while their influences on distinct functions remain understudied. Here, we linked 753 treehole bacterial communities to two distinct types of EFs, including the three broad functions of respiration, metabolic activity and cell yield and the four narrow functions related to specific organic matter degradation. Towards high occurrence of phylotypes or traits, the dependency of broad EFs on species abundance or traits increased, whereas the dependency of narrow functions decreased. Among the immense number of bacterial phylotypes, the relative abundance of only 5.05% of phylotypes (that is, 542 phylotypes), but accounting for 68.60% of total abundance, were significantly related to both distinct EFs ranging from 2 to 7 functions, the level of which was used to quantify species functional generality. Such 'low species number, high relative abundance and strong functional generality' features for these 542 phylotypes could be further potentially linked to their enriched functional genes involved in cellular processes including nutrient acquisition, environmental adaptation and cell growth. Our study highlights the key role of a handful of microbial species in determining and anticipating distinct EFs by explicitly considering their abundance and trait attributes.

Introduction

One of the most important questions in microbial ecology is how to resolve the relationship between community structure and ecosystem functions (EFs) (Vitousek and Hooper, 1994). The community-level diversity metrics including species richness and community composition are proposed to affect EFs such as the bacterial species composition for community respiration (Bell *et al*., 2005) and the denitrifier community for the rates of denitrification and N_2O production (Cavigelli and Robertson, 2000). In addition, the species-level attributes such as species abundance, identity and traits also play important roles in determining EFs. A skewed species abundance distribution shows that a few of species are very abundant whereas a large number are represented by low-abundance individuals, and the pattern is commonly observed in microbial communities across various habitats (Nemergut *et al*., 2011; Pedrós-Alió, 2012; Lynch and Neufeld, 2015). Such a pattern indicates that not all species contributes equally to the variations in EFs (Banerjee *et al*., 2018), and species abundance could be associated with their functional role in microbial communities. For example, the abundant and rare species in treehole bacterial communities are involved in two fundamentally different types of EFs, that is, the broad (e.g., respiration and ATP production) and narrow (e.g., the degradation of complex substrates) functions respectively (Rivett and Bell, 2018). Further, species identity and traits, rather than species richness, are also pivotal for narrow EFs such as the degradation rate of chitin being determined by the presence or absence of Agrobacterium-related species (Peter *et al*., 2011). However, the quantitative relationships between EFs and the species attributes, including species abundance or traits, remains understudied, especially regarding these distinct types of EFs. Specifically, to what extent do these distinct EFs depend on species attributes, and are there predictable patterns in such a dependency along the gradient of the occurrence of species attributes? Are there differentiated performances among species in terms of their contributions to EFs? If so, what is the genomic basis in explaining the differentiated performances?

To answer these questions, we reanalysed the comprehensive dataset from Rivett and Bell (2018) consisting of 753 bacterial communities from rainwater-filled treeholes and two distinct EFs associated with leaf litter degradation (Fig. 1 and Supporting Information). Briefly,

Received 8 February, 2021; accepted 11 June, 2021. *For correspondence. E-mail jjwang@niglas.ac.cn; Tel. +86-25-86882219; Fax. +86-25-86882219.

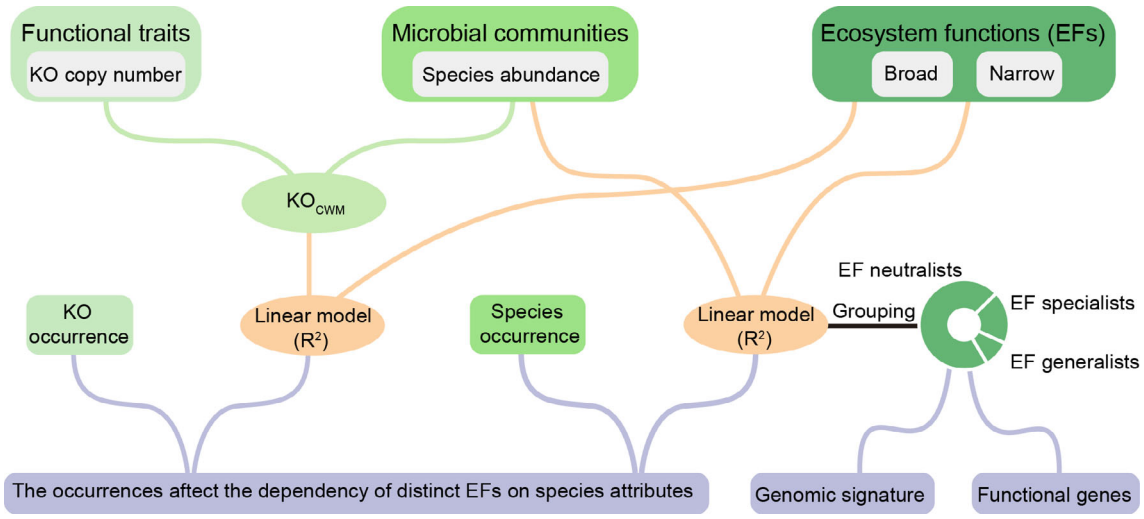

Fig. 1. The framework of the experimental design. We explored 753 bacterial communities with two fundamentally different types of ecosystem functions, including the broad (cell respiration, cell yields and ATP production) and narrow functions (the activities of four enzymes associated with organic matter degradation, Table S1). The study has three main aims in: (i) determining the relationships between the occurrence of species attributes and the dependence of different types of ecosystem functions on them, (ii) identifying the differentiated performances among phylotypes in terms of their contributions to distinct EFs and (iii) exploring the genomic basis in explaining the differentiated performances. EFs, ecosystem functions; KO, Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology. KO_{CWM}, the community-level weighted means of KO. [Color figure can be viewed at wileyonlinelibrary.com]

we obtained 10,729 phylotypes at the 97% 16S rRNA gene sequence similarity level. We further retrieved functional traits for 5060 phylotypes regarding genomic signatures and functional genes of Kyoto Encyclopedia of Genes and Genomes (KEGG) orthology (KO) by mapping the representative sequence against the prokaryotic genomes in the RefSeq database (O'Leary *et al*., 2015). The community-level weighted means (CWM) of each KO (that is, KO_{CWM}) was calculated for each sample as the mean of trait values present in the community weighted by the relative abundance of the corresponding phylotypes (Garnier *et al*., 2004). For EFs, there were seven measures grouped into the broad and narrow EFs according to Rivett and Bell (2018): cell respiration, cell yields and metabolic potential were considered as broad EFs; the activities of four enzymes associated with organic matter degradation were considered as narrow EFs, including β-glucosidase (breaks down cellulose), β-chitinase (breaks down chitin), phosphatase (breaks down organic phosphates) and xylosidase (cleaves xylose, a component of hemicelluloses; Table S1). EFs could be viewed as a continuum of functions from relatively narrow to relatively broad, and those located towards the two ends of the spectrum are selected to capture the distinct or contrasting features in microbes associated with fundamentally different types of EFs. Our selection of EFs is consistent with the classification of EF, namely, the narrow EFs refer to the processes that involve a specific physiological pathway, such as litter decomposition, which are performed by a phylogenetically constrained group of organisms, whereas the broad EFs refer to the processes that involve multiple distinct steps, such as soil respiration, which are carried out by a wide range of organisms (Schimel and Schaeffer, 2012; Fierer, 2017).

Results and discussion

These EFs showed significant associations with the relative abundance of individual phylotype and also with KO_{CWM} ($P < 0.05$, *F*-test, Fig. 2). We interpreted such associations as the dependency of EFs on species abundance or KO_{CWM} and quantified them with the coefficient of linear regression (referred to as effect size hereafter). KO_{CWM} had a larger effect size than species abundance for broad EFs, but had a smaller effect size for narrow EFs (Fig. 2C). Our results not only emphasized the importance of species abundance for EFs, being consistent with previous findings (Rivett and Bell, 2018), but also revealed that the traits had stronger association with broad but not narrow functions when compared with species abundance.

Further, we found that the dependency of EFs on species abundance or KO_{CWM} was related to the occurrence of species attributes, and such a relationship showed contrasting patterns for two distinct types of functions (Fig. 2A and B). For instance, the effect size of species abundance or KO_{CWM} for broad EFs significantly ($P < 0.01$, Spearman rank test) increased towards higher occurrence of phylotypes or KOs (upper panels in

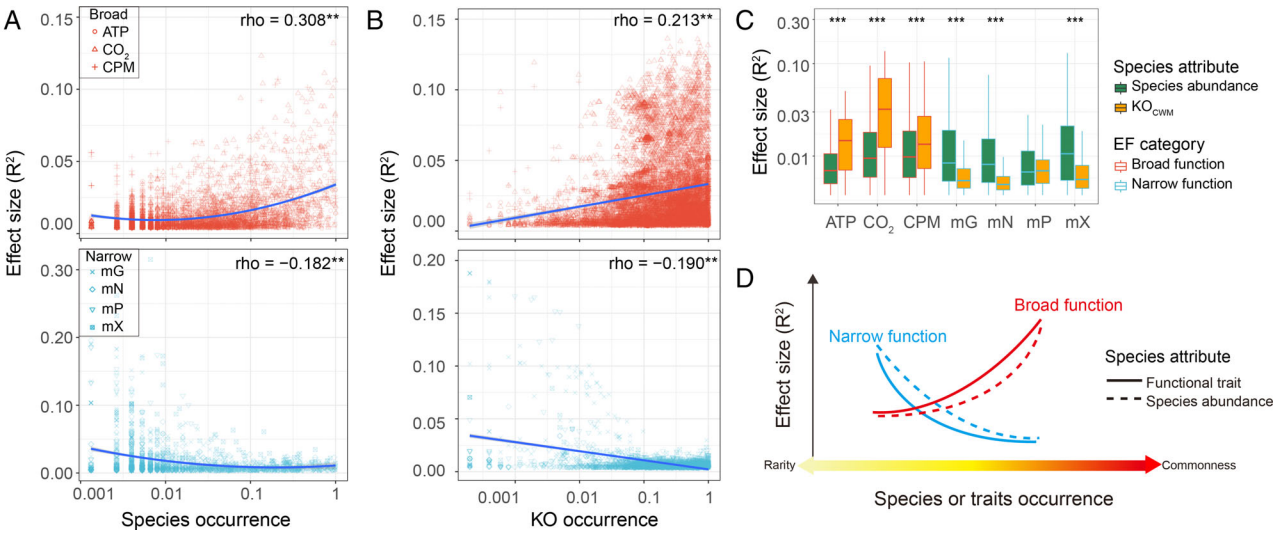

Fig. 2. The dependency of ecosystem functions on species abundance and traits was related to the occurrence of phylotypes and traits. The linear regression model was used to fit the relationships between species abundance or traits (KO_{CWM}) and the EFs (for details about the model, see the Supporting Information). Such relationships were interpreted as the dependency of a specific function on species abundance or KO_{CWM} and quantified with the correlation coefficient of the regression (R^2, hereafter referred to as effect size). These effect sizes were then plotted against the occurrence of phylotypes (A) or KO (B). The effect sizes of species abundance and KO_{CWM} for different EFs varied significantly (Wilcoxon test) (C). *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. Based on these analyses, we conceptually summarized the relationships between EFs and the phylotypes in terms of the species and trait occurrences (D). [Color figure can be viewed at wileyonlinelibrary.com]

Fig. 2A and B, and Figs. S1 and S2). For narrow functions (lower panels in Fig. 2A and B), the effect size of species abundance or KO_{CWM} showed significant ($P < 0.01$, Spearman rank test) decreasing trends with increasing occurrence of phylotypes or KOs. Such patterns were further supported by the same analyses performed on a subset of 1000 phylotypes (100 permutations), which consistently showed significant associations between the different types of EFs and species attributes (Fig. S3).

The contrasting patterns indicate that different types of EFs are closely related to the occurrence of species attributes, which agrees with previous findings regarding species abundance and EFs (Fuhrman, 2009; Jousset *et al*., 2017). For example, the abundant bacterial groups contribute mainly to broad functions, such as bacterial biomass production (Cottrell and David, 2003) and the flux of dissolved organic matter being predominantly affected by an abundant marine bacteria group SAR11 (Malmstrom *et al*., 2005). However, the rare phylotypes have been proposed to play an important role in determining EFs, including the degradation rates of chitin and cellulose in a manipulative experiment of aquatic bacterioplankton (Peter *et al*., 2011), the activities of four extracellular enzymes in soil (Chen *et al*., 2020), and the stability of crop mycobiomes (Xiong *et al*., 2021). Such effects of rare phylotypes on EFs may be explained by three possible mechanisms: stronger phylotypes activity, increased functional diversity and beneficial influences on the abundant phylotypes (Jousset *et al*., 2017).

These findings could be summarized by a framework that illustrates the relationships between species attributes and EFs in terms of the occurrence of species and their traits (Fig. 2D): the dependency of broad EFs on species abundance or KO_{CWM} is positively related to the occurrence of species or functional traits, whereas the dependency of narrow EFs increases towards high rarity of species or functional traits. Species with large occurrences likely adapt to a wide range of environments and enable the performance of broad functions, such as respiration or ATP production, whereas rare species may specialize in particular environments with their phylogenetically conserved features, such as limiting resource or strong competition through the degradation of specific compounds. Thus, not all species contribute equally to a single function (Banerjee *et al*., 2018), which is partly due to its specific trade-offs in the ability to perform different functions (Peter *et al*., 2011).

Based on the associations between species abundance and EFs, we found that some species showed significant linear regressions with more than one function (Figs. 3A and S4). Such a capability was defined as species functional generality, the level of which was quantified as the cumulative number of EFs showing significant ($P < 0.05$, F statistics) relations with species abundance. When both types of functions were considered (Fig. S4),

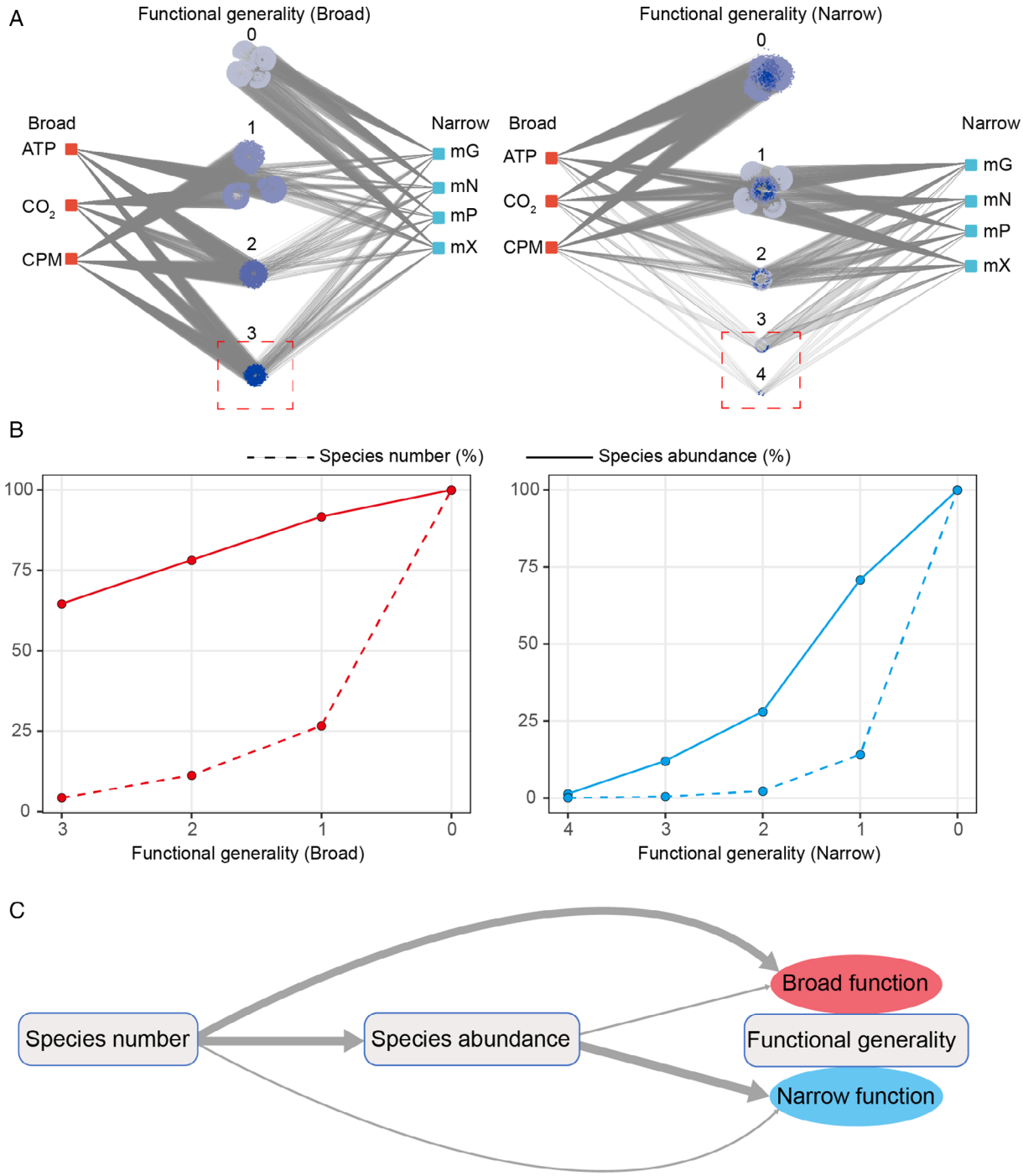

Fig. 3. The species functional generality for two types of ecosystem functions and their relationships with species abundance and species number. Species functional generality was quantified as the cumulative number of the functions showing significant ($P < 0.05$) relationships with the abundance of a phylotype. For each phylotypes, functional generality for both types of EFs were linked (A) and compared (B). The associations between phylotypes (blue points) and both types of functions (the red and green squares) were illustrated by the network where the edges represent the significant associations. In the left panel of (A), the phylotypes showing the same functional generality (the digits) for broad functions were illustrated with the same magnitude of blue colour, whereas in the right panel of (A), the phylotypes with the same colour schemes were rearranged by the functional generality for the narrow functions. An alluvial plot illustrating these associations was shown in Fig. S4. In (B), the curves represent the change of species relative abundance and their number along functional generality for each specific EF types. A conceptual diagram about the relationships was shown in (C). A detailed explanation of the diagram was provided in the main text. [Color figure can be viewed at wileyonlinelibrary.com]

0.53% of phylotypes (i.e., 57 phylotypes) showed a functional generality score over five, and 1.96% (i.e., 210 phylotypes) had a generality score over four, while 64.20% showed zero functional generality. For the broad EFs, 4.27% of phylotypes (i.e., 458 phylotypes) but accounting for 64.60% of total abundance, were associated with the highest levels of functional generality (that is, $= 3$) (Fig. 3A and B). Such a high functional generality for broad functions is likely to be consistent with previous literatures showing that approximately 20% of species accounting for 80% of the total community abundance are responsible for 80% of the metabolic energy flux of the ecosystem (Dejonghe *et al*., 2001; De Vrieze and Verstraete, 2016). For the narrow EFs, fewer phylotypes (0.45%) but with a relatively larger abundance showed high functional generality values when compared with broad functions (Fig. 3B). For example, the average abundance of the phylotypes with functional generality (≥ 1) for narrow EFs was up to four times higher than that for broad EFs (Table S2). Specifically, only 48 phylotypes but accounting for 12.10% of total abundance were associated with the highest functional generality (that is, ≥ 3) for narrow EFs (Fig. 3A and B). Such distinct patterns indicate that the functional generality for narrow functions depends on more species abundance than species number, whereas species number has a greater effect on the functional generality for broad EFs (Fig. 3C). The high functional generality for narrow EFs suggests that the growth of these species is closely associated with multiple narrow functions, including their participation in the functions or their utilization of intermediate substrates. Few species can perform multiple narrow functions simultaneously due to great metabolic requirements. Therefore, it is very likely that the resource, derived from the degradation of specific complex substrates, provides great advantages to their growth.

Moreover, we found that a few species showed high predictive power for EFs regardless of the type of functions (that is, broad or narrow EF). All phylotypes could be further classified into four groups based on their relationships with the two types of EFs, that is: (i and ii) EF specialists (EFS) significantly associated with broad or narrow functions (that is, EFS-B and EFS-N respectively), (iii) EF generalists significantly associated with both types of functions and finally (iv) EF neutralists showing non-significant relations with all functions (Fig. S5). As expected, these groups exhibited different performances in predicting EFs when using a random forest model (Breiman, 2001) with species abundance or species richness as explanatory variables. For instance, regarding species abundance or KO_{CWM}, the EF specialists for both types of functions showed better predictions of the broad and narrow EFs respectively (Fig. 4A and B). Interestingly, species richness of EFS-N better predicted the narrow than broad functions, particularly for the cellulose utilization rate with an average coefficient of 0.4 (Fig. 4C). Unexpectedly, the EF generalists were generally better at predicting both types of EFs than the EF specialists with respect to their species abundance or KO_{CWM} (Fig. 4A and B). Notably, the EF generalists consisted of a relatively small fraction of phylotypes (i.e., 5.05%, 542 phylotypes) but had the greatest relative abundance of 68.60% (Fig. S5B). A similar phenomenon is observed for soil bacteria, in which 2% of species account for nearly half communities worldwide (Delgado-Baquerizo *et al*., 2018). The characteristics of the generalists could be summarized as 'low species number, high relative abundance and strong functional generality'. This indicates that a few species but with high abundance are associated with EFs, however, the genomic basis underlying these patterns are understudied.

We further performed comparative genomic analysis among the above three species groups (that is, the EF generalist and two EF specialist groups) to explore the differences in their genomic signatures and genes enriched or depleted in cellular processes. The genomic signatures like genome size, the number of protein-coding genes and GC content were significantly higher in EF generalists than those in the other two groups ($P < 0.05$, Wilcoxon test; Fig. S6A). Larger-genome EF generalists are usually enriched in the regulation and secondary metabolism genes which help the organisms to take up nutrients and cope with environmental stress (Konstantinidis and Tiedje, 2004); thus, these generalists are likely to have larger environmental range sizes (i.e., the breadth of habitats) than other organisms, resulting in greater ubiquity as observed in soil bacteria (Barberán *et al*., 2014). For functional genes, the Shannon index of KOs was significantly higher for the EF generalists than for the specialists ($P < 0.05$, Wilcoxon test, Fig. S6A). Furthermore, regarding the higher-level KEGG pathways (Kanehisa *et al*., 2016), there were significant differences ($P < 0.05$, Wilcoxon test) among the three species groups in KO copy numbers of numerous functional pathways (Fig. 4D), which were summarized into three categories: efficient nutrient acquisition, diverse habitat adaptation and fast growth strategy.

i. Efficient nutrient acquisition: For example, compared with the EF specialists, the generalists were significantly enriched in genes involved in flagella assembly and ATP binding cassette transporter ($P < 0.05$, Wilcoxon test; Fig. S7). Such enriched genes relevant to bacterial motility and substrate-specific transport allow generalists to take up the transient nutrient in the environment (Smriga *et al*., 2016). Among the generalists, there were also significantly enriched gene responsible for the degradation of fatty acids ($P < 0.05$;

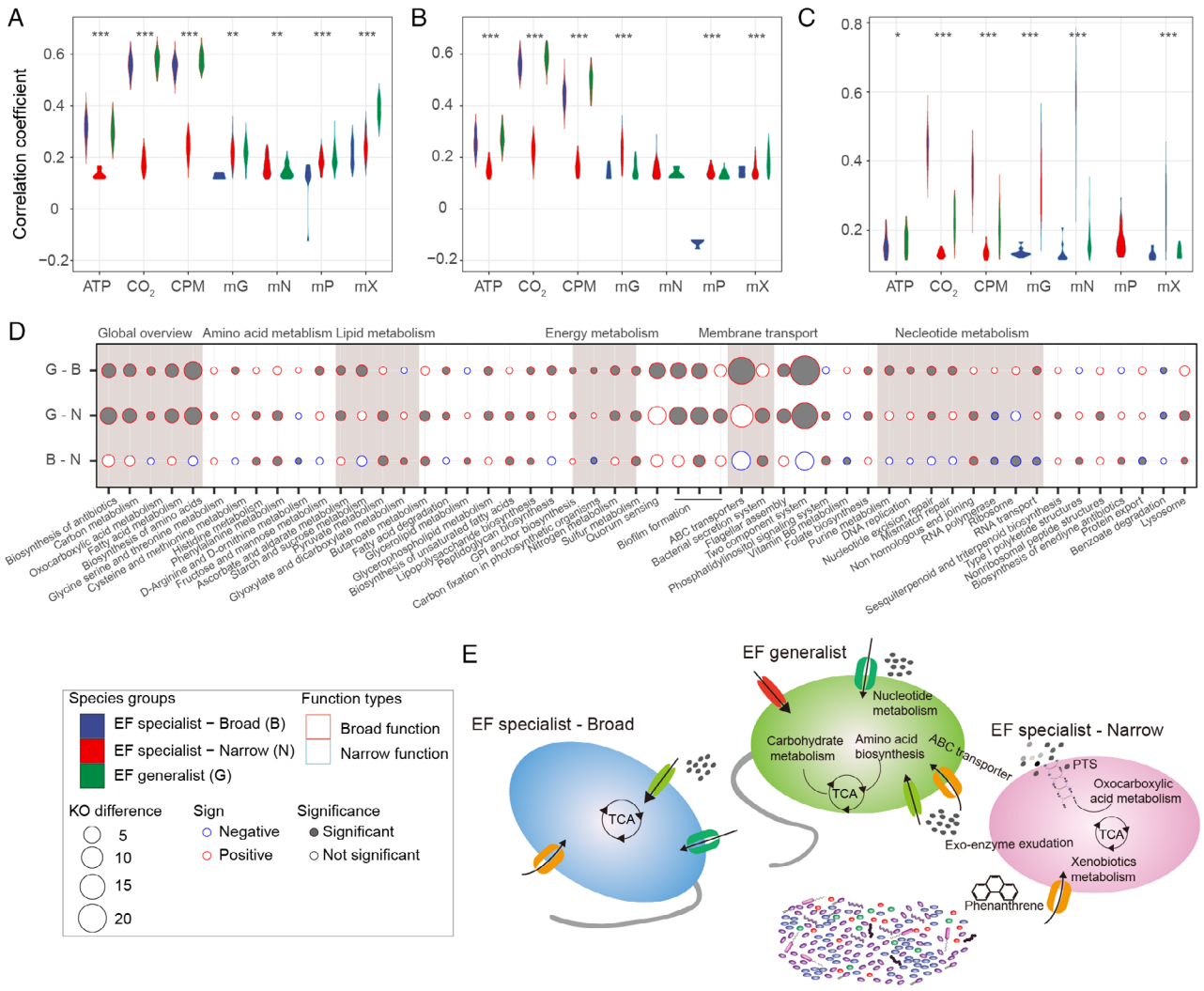

Fig. 4. Prediction of ecosystem functions using the attributes of four species groups and their functional difference. These four groups were EF specialists significantly associated with broad (i) or narrow (ii) functions, (iii) EF generalists and finally (iv) EF neutralists (Fig. S5). The ecosystem functions were predicted using the attributes of each group, including species abundance (A), KO_{CWM} (B) and species richness (C) with a random forest model. The predictive power of the model regarding these attributes was quantified with the Spearman rank correlation coefficient between the predicted and observed values for each function. The criterion for species classification and the procedure for calculating the correlation coefficient were detailed in the Supporting Information. The above analysis was repeated 100 times, and only the significant coefficients were illustrated. The differences in correlation coefficients between groups was tested with Wilcoxon test. *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. The pathways with KO copy number showing a significant difference between any pair of species groups (Wilcoxon test) were shown (D). All functional differences were conceptually summarized in (E). For better visualization, the EF neutralist species were not shown. [Color figure can be viewed at wileyonlinelibrary.com]

Fig. 4D) which are essential components of membranes and an important source of metabolic energy for all organisms (Fujita *et al*., 2007).

ii. Diverse habitat adaptation: Compared with the EF specialists, the generalists showed a higher abundance of the genes responsible for biofilm formation (Figs. 3D and S9), which concurs with the most abundant genus *Pseudomonas* (Fig. S6B) generally capable of forming biofilms (Masák *et al*., 2014). Biofilm formation may provide microbes with numerous ecological advantages, including the resistance to environmental disturbance, metabolic cooperation among species in close proximity and the acquisition of novel genetic traits from the surroundings (Davey and O'toole, 2000). Among the EF specialists, the EFS-N species harboured more genes involved in xenobiotics metabolism, such as benzoate degradation (Figs. 4D and S7), which agrees with the top two taxa of *Sphingomonas* and *Arthrobacter* genera (collectively accounting for ~52.50% of total abundance,

Fig. S6B) known to have a wide range of xenobiotic-biodegradative abilities (Vandera *et al*., 2015; Ravintheran *et al*., 2019). These xenobiotic-related genes potentially enhance the adaptation of these specialists to habitats containing toxic compounds.

iii. Fast growth strategy: Between the EF generalists and specialists, we found a non-significant difference in the number of 16S rRNA genes, an estimate for rRNA operon copy number (Fig. S6A), which is usually linked to the maximum growth rate of a bacterium (Roller *et al*., 2016). However, the generalists showed significantly enriched genes involved in nucleotide metabolism, DNA replication and repair and amino acid biosynthesis (Fig. 4D). The associations with these genes concur with the scenario in which the generalists use a fast growth strategy requiring more energy allocated to the manufacture of amino acids and the regulation for DNA biosynthesis (Molenaar *et al*., 2009).

In conclusion, our results showed that only a few species but dominating the bacterial communities were significantly associated with the two fundamentally different types of EFs, and their relative abundance and traits well predicted EFs. These features could be summarized as 'low species number, high relative abundance and strong functional generality'. Such distinctive performance of these species may be relevant to the predominance of their functional genes associated with efficient nutrient acquisition, diverse habitat adaptation and fast growth strategy. Here our study sheds light on the importance of a relatively few dominant microbial species in the prediction of distinct EFs based on species attributes, and provides the understanding of their characteristics leveraging available genomic data.

Experimental procedures

Sequence analyses

The 16S rRNA gene sequence for 753 aquatic samples was downloaded from NCBI SRA database (SRR7136127–SRR7136875 under the accession number PRJNA453972). The V4 region of 16S rRNA gene was amplified with using the primers 515f/806r and then sequenced on the Illumina MiSeq (250-bp paired end) platform. It should be noted that we did not consider fungal communities as fungal activity was eliminated by supplementing fungicide (cycloheximide) during experimental incubations. The bacterial dataset was re-analysed using our custom pipeline. Briefly, FastQC v0.11.8 was used to check the Phred quality of the sequence and the frequency of potentially contaminated adapters (Andrews, 2010). The sequence with the average Phred quality lower than 25 within a 4-bp sliding window, was then trimmed using the paired end mode of Trimmomatic v0.39 and the resulting reads shorter than 250 bp were discarded (Bolger *et al*., 2014). Due to a large amount of memory required by QIIME 2 program (Bolyen *et al*., 2019), the clustering and selection of operational taxonomic units and taxonomy assignment were achieved using the script 'pick_open_reference_otus.py' in QIIME v1.9.1 (Caporaso *et al*., 2010). In QIIME1, uclust (Edgar, 2010), a high-speed and low memory usage clustering method, was used to cluster the sequence with the similarity threshold being 97%, a frequently used cutoff to group microbes into species (Yarza *et al*., 2014), which we refer to as 'phylotypes' in the text. For each phylotype, taxonomy assignment was also performed using uclust, which aligns a query sequence against Greengene database v13.8 (DeSantis *et al*., 2006), then retrieves at least three hits from the database for the query, and finally assigns the most specific taxonomic label of the hit to the query. The phylotypes were rarefied with the minimum sequence abundance of all samples (5660 sequence) using the '*rrarefy*' method in VEGAN package (Dixon, 2003). To avoid the potential sequencing error and the bias caused by the uneven sequencing efforts, the phylotypes with the total abundance across all samples less than three and the number of occurring samples less than two were discarded, resulting in 10,726 phylotypes from all samples. A relatively stringent filter criteria (at least 100 individual across all samples and occurred only in at least 10 samples) mentioned in the original paper (Rivett and Bell, 2018) was not applied here in order to capture more phylotypes in communities. In addition, we perform additional sequence analysis using 100% similarity threshold and found that Shannon diversity estimated from 97% showed strong correlation with the index based on 100% (Spearman rho $= 0.90$, $P = 0$, Fig. S8), indicating marginal effects of similarity threshold on the estimation of species diversity.

Functional annotation

We mapped the phylotypes with the corresponding microbial organism whose genome sequence is available through the alignment of their 16S rRNA genes. Although the single marker gene alignment-based method is controversial in the term of accuracy, similar methods have been widely used (Barberán *et al*., 2014). In brief, bacterial and archaeal genome files (21,788 species with unique taxa identity) in GBK format were downloaded from the NCBI RefSeq database (date: April 2019). For each organism, the 16S rRNA gene greater than 1000 bp was extracted and pooled together as the reference database. When multiple copies of 16S rRNA genes are

present, the longest one was selected as the representative for the species. The sequence of phylotypes was aligned against the above reference database using BLASTN program with the e-value being '1e-5'. The hit with sequence identity larger than 97% and the proportion of the alignment length larger than 80% of the query length was considered as the closely related representative of the query phylotypes. Among all 10,729 phylotypes, 5060 were successfully linked to 1282 unique reference genomes (Table S3), whereas the remaining 5669 phylotypes had no alignments in the NCBI prokaryote RefSeq database or the alignments that could not meet the stringent criteria. No closely related genomic sequence was retrieved for more than half of the phylotypes, partly because a high proportion of bacteria and archaea across most biomes remains uncultured and not represented in the database (Steen *et al*., 2019). For the phylotypes ($N = 5060$) with matched genome sequence, several genomic signatures, including genome size, GC content, the number of protein-coding genes, total genes and 16S rRNA gene were calculated. To infer metabolic potential for these phylotypes, the amino acid sequence of the protein-coding genes in each genome were aligned against the orthologue in the Kyoto Encyclopedia of Genes and Genomes (KEGG) database (Kanehisa *et al*., 2016) (date: 2017-09) using DIAMOND v0.8.22 (Buchfink *et al*., 2015) with e-value of 1e-5. The top 10 hits in the alignment were considered as the candidates, and the consensus KO was then determined as the final annotation. When more than one KOs present among these hits, the consensus KO was assigned to the query by the majority rule. According to the links between KO and KEGG pathway (Kanehisa *et al*., 2016), the abundance of each pathway was calculated as the total number of KOs present in the pathway for each species. Unless stated explicitly, default parameters are used in all programs abovementioned.

The functional prediction approaches have been controversially discussed regarding the representativeness of real functional potential of microbial community. We considered that our approach is valid and valuable, especially when genome sequences for most microbes are hardly accessible by cultivation technology currently available, and provided five reasonings as below: (i) The functional prediction approaches with genome-mapping are extensively used and similar principles have been implemented in numerous tools such as PICRUST (Langille *et al*., 2013; Douglas *et al*., 2020). These tools are still under heavy development with very recent updates, and very popularly referred in literature. (ii) Compared with the available tools, we applied a relatively more conserved mapping approach with a stringent criterion (> 97% sequence identity and > 80% alignment length), but without any extrapolations of functions for other unaligned phylotypes. For instance, PICRUST depends on the phylogenetic proximity between reference genome and environmental strains to predict functional composition for all phylotypes (Langille *et al*., 2013). (iii) The validity of genome-mapping approach is well supported by previous studies. For example, through matching 16S rRNA sequences against the RDP database derived from bacterial genomes, soil bacteria with larger genomes and more metabolic versatility are more likely to have larger environmental and geographical distributions (Barberán *et al*., 2014). (iv) The subset of phylotypes with available genomes in our study well represented the whole microbial community. We found that 5060 phylotypes with available genomes accounted for 96.1% of the total abundance, and their Shannon diversity strongly correlated with that of all phylotypes (Fig. S9). (v) The genomic features considered in this study, such as genomic size, the count of CDS and GC content, are not available in other programs like PICRUST.

Statistical analyses

Three datasets were used in statistical analyses: (i) 10,729 phylotypes from 753 samples after filtering (see Section Sequence analyses); (ii) the 5060 phylotypes with genome sequence available, the corresponding genomic signatures and the KO copy number (see Section Functional annotation); (iii) seven functional measurements across all 753 samples from the original Supporting Information (Rivett and Bell, 2018). The overview of statistical analyses was summarized in Fig. 1.

(i) *The relationships between functional measures and species abundance or genomic traits across samples*.

Two **species attributes** are focused in the study, namely species abundance and genomic trait represented as KOs. To test the relationship between genomic traits and measured functions, the community weighted level mean values of KOs (KO_{CWM}) is calculated for each sample using the '*functcomp'* method in FD package (Laliberté and Legendre, 2010). **A linear regression model** was used to fit the relationship between each functional measure and species abundance or KO_{CWM} in all samples. For species abundance, the formula $y = b_1 log_{10}(x+1) + b_0$ was applied, whereas another formula $y = b_1 x + b_0$ was used for KO_{CWM}, in which y is a functional measure and x is the relative abundance of phylotypes or KO_{CWM} of samples. The coefficient (R^2) of the linear regression were used to represent the dependence of EF on species abundance or genomic trait.

(ii) *The relationships between coefficient and the occurrence of species or trait*.

For each phylotype and KO gene family, the R^2 of linear regression for different functional measures was

calculated and referred to as the effect size. Then, species occurrence was defined as the number of samples containing the phylotype divided by the number of samples. KO occurrence was defined as the total number of times KO present in phylotypes divided by the total phylotype number. Linear regression and quadratic regression are selected using the Akaike information criterion (AIC) to fit the relationship between the effect size and the occurrence of phylotypes or traits.

In addition, a resampling analysis was performed to test the consistency of the correlation between the effect size and species occurrence, or the effect size and KO occurrence. Firstly, a subset of 1000 phylotypes was randomly selected from phylotypes in all samples. For species abundance, the corresponding effect size of these phylotypes and their occurrence were calculated. For genomic traits, KO_{CWM} was calculated using the abundance and KO copy number of the randomly selected 1000 phylotypes, rather than all species. Secondly, the Spearman's rho correlation between the effect size and the occurrence of species and traits was then calculated respectively. Finally, these procedures were replicated 100 times and estimated the distribution of Spearman's rho correlation.

Species occurrence was strongly correlated with their mean relative abundance across samples (Fig. S10), as expected. The relationship between its occurrence (presence or absence data) and EFs across samples cannot be quantified due to the inconsistency of variable types (0/1 and continuous respectively). Instead, we evaluated the correlations between relative species abundance and fundamentally different EFs across samples using the linear regression model and found that towards high species occurrence, the dependency of EFs on species abundance exhibited contrasting trends regarding different types of EFs. These results highlight the differentiating roles of species along the occurrence gradient in explaining the variation of EF.

(iii) *The network illustration of the association between species and ecosystem functioning, and the classification of species*.

The associations between species abundance and all EFs in the study were illustrated with an undirected network implemented in Cytoscape v3.8.0 (Shannon *et al*., 2003). The nodes in the network are phylotypes and seven functions, and edges between phylotypes and functions are significant association between species abundance and both types of functions. Based on the significance of these associations and function category, the phylotypes were classified into four groups: (i and ii) EF specialists (EFS) significantly associating with broad or narrow functions (that is, EFS-B and EFS-N respectively), (iii) EF generalists significantly associating with both functions, and finally (iv) EF neutralists showing non-significant relations with any function.

(iv) *Prediction of EF based on classified species*.

The random forest (RF) regression model (Breiman, 2001) was used to predict different types of EFs. The RF model could incorporate a large number of predictors, like the abundance of numerous phylotypes or genomic traits in our study. The procedure of constructing the model was described as follows: (i) 60% of the samples was randomly selected as the training dataset and the remaining 40% as the test dataset. (ii) The RF model for each function was generated using species abundance or KO_{CWM} for each of four species groups based on the training dataset. When the number of predictors (phylotypes or KOs) in groups is large, the maximum of 1000 items were randomly selected from the group as the representative predictors in each run to save the running time. (iii) The fitted RF model was used to predict the EF using the test dataset; and (iv) the spearman rank correlation coefficient between the observed and predicted values of EF was calculated. These four steps were replicated 100 times to evaluate the distribution of the correlation coefficient. The RF algorithm implemented in the randomForest package was applied to construct the model (Liaw and Wiener, 2002). All reported results about the RF model were trained with 1000 trees. Besides species abundance or KO_{CWM}, species richness of these four species groups was also utilized to predict EFs using the same procedure. Species richness were calculated using the methods in the VEGAN package (Dixon, 2003).

(v) *Difference in genomic signatures and functional pathway between species groups*.

When the corresponding genome sequence for the representative phylotypes was retrieved using the single-marker gene mapping method, a common case occurs that more than one species share the same genome. However, our stringent criteria (≥ 97% similarity and ≥ 80% alignment length of 16S rRNA gene sequence) ensures that the matched genome sequence is a reliable proxy for the representative species. Thus, the replicate genome was not excluded in the downstream analysis.

For the phylotypes with genome available, genomic signatures, including genome size, GC content, CDS count, all gene count and 16S rRNA gene count were estimated (see Section Sequence analyses). The Wilcoxon test for two independent samples was used to compare the difference of mean values in each genomic signature between any two groups among the three. KEGG pathway with significantly higher number of KOs in each group are screened out in a similar way. Briefly, for each phylotype, the total KO copy number in each pathway at the lowest level were counted as the relative abundance of the pathway according to the hierarchical organization in KEGG pathway (Kanehisa *et al*., 2016). Then, for each pathway, the difference in the relative

abundance between any two groups among the three was checked using the Wilcoxon test at both sides. Then pathways with the *P* values adjusted by the Holm-Bonferroni method less than 0.05 were discussed in the study.

Acknowledgements

This study was supported by National Natural Science Foundation of China (91851117), CAS Key Research Program of Frontier Sciences (QYZDB-SSW-DQC043), CAS Strategic Pilot Science and Technology (XDA20050101), National Natural Science Foundation of China (41871048, 42002304, 41571058), Nanjing institute of Geography and Limnology, Chinese Academy of Science starting grant (Y9SL031). We also thank Thomas Bell for kindly checking and sharing of amplicon sequence data, and for further discussion.

Author contributions

The research was conceived by Jianjun Wang. Data analyses were undertaken by Minglei Ren with the contribution of Jianjun Wang. The first draft of the manuscript was written by Minglei Ren. All authors discussed and revised the manuscript.

References

Andrews, S. (2010) FastQC: A quality control tool for high throughput sequence data. URL https://www.bioinformatics.babraham.ac.uk/projects/fastqc/.

Banerjee, S., Schlaeppi, K., and van der Heijden, M.G.A. (2018) Keystone taxa as drivers of microbiome structure and functioning. *Nat Rev Microbiol* **16**: 567–576.

Barberán, A., Ramirez, K.S., Leff, J.W., Bradford, M.A., Wall, D.H., and Fierer, N. (2014) Why are some microbes more ubiquitous than others? Predicting the habitat breadth of soil bacteria. *Ecol Lett* **17**: 794–802.

Bell, T., Newman, J.A., Silverman, B.W., Turner, S.L., and Lilley, A.K. (2005) The contribution of species richness and composition to bacterial services. *Nature* **436**: 1157–1160.

Bolger, A.M., Lohse, M., and Usadel, B. (2014) Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* **30**: 2114–2120.

Bolyen, E., Rideout, J.R., Dillon, M.R., Bokulich, N.A., Abnet, C.C., Al-Ghalith, G.A., *et al*. (2019) Reproducible, interactive, scalable and extensible microbiome data science using QIIME 2. *Nat Biotechnol* **37**: 852–857.

Breiman, L. (2001) Random forests. *Machine Learn* **45**: 5–32.

Buchfink, B., Xie, C., and Huson, D.H. (2015) Fast and sensitive protein alignment using DIAMOND. *Nat Methods* **12**: 59–60.

Caporaso, J.G., Kuczynski, J., Stombaugh, J., Bittinger, K., Bushman, F.D., Costello, E.K., *et al*. (2010) QIIME allows analysis of high-throughput community sequencing data. *Nat Methods* **7**: 335–336.

Cavigelli, M.A., and Robertson, G.P. (2000) The functional significance of denitrifier community composition in a terrestrial ecosystem. *Ecology* **81**: 1402–1414.

Chen, Q.-L., Ding, J., Zhu, D., Hu, H.-W., Delgado-Baquerizo, M., Ma, Y.-B., *et al*. (2020) Rare microbial taxa as the major drivers of ecosystem multifunctionality in long-term fertilized soils. *Soil Biol Biochem* **141**: 107686.

Cottrell, M.T., and David, K.L. (2003) Contribution of major bacterial groups to bacterial biomass production (thymidine and leucine incorporation) in the Delaware estuary. *Limnol Oceanogr* **48**: 168–178.

Davey, M.E., and O'toole, G.A. (2000) Microbial biofilms: from ecology to molecular genetics. *Microbiol Mol Biol Rev* **64**: 847–867.

De Vrieze, J., and Verstraete, W. (2016) Perspectives for microbial community composition in anaerobic digestion: from abundance and activity to connectivity. *Environ Microbiol* **18**: 2797–2809.

Dejonghe, W., Boon, N., Seghers, D., Top, E.M., and Verstraete, W. (2001) Bioaugmentation of soils by increasing microbial richness: missing links. *Environ Microbiol* **3**: 649–657.

Delgado-Baquerizo, M., Oliverio, A.M., Brewer, T.E., Benavent-González, A., Eldridge, D.J., Bardgett, R.D., *et al*. (2018) A global atlas of the dominant bacteria found in soil. *Science* **359**: 320–325.

DeSantis, T.Z., Hugenholtz, P., Larsen, N., Rojas, M., Brodie, E.L., Keller, K., *et al*. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol* **72**: 5069–5072.

Dixon, P. (2003) VEGAN, a package of R functions for community ecology. *J Veg Sci* **14**: 927–930.

Douglas, G.M., Maffei, V.J., Zaneveld, J.R., Yurgel, S.N., Brown, J.R., Taylor, C.M., *et al*. (2020) PICRUSt2 for prediction of metagenome functions. *Nat Biotechnol* **38**: 685–688.

Edgar, R.C. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**: 2460–2461.

Fierer, N. (2017) Embracing the unknown: disentangling the complexities of the soil microbiome. *Nat Rev Microbiol* **15**: 579–590.

Fuhrman, J.A. (2009) Microbial community structure and its functional implications. *Nature* **459**: 193–199.

Fujita, Y., Matsuoka, H., and Hirooka, K. (2007) Regulation of fatty acid metabolism in bacteria. *Mol Microbiol* **66**: 829–839.

Garnier, E., Cortez, J., Billès, G., Navas, M.-L., Roumet, C., Debussche, M., *et al*. (2004) Plant functional markers capture ecosystem properties during secondary succession. *Ecology* **85**: 2630–2637.

Jousset, A., Bienhold, C., Chatzinotas, A., Gallien, L., Gobet, A., Kurm, V., *et al*. (2017) Where less may be more: how the rare biosphere pulls ecosystems strings. *ISME J* **11**: 853–862.

Kanehisa, M., Furumichi, M., Tanabe, M., Sato, Y., and Morishima, K. (2016) KEGG: new perspectives on genomes, pathways, diseases and drugs. *Nucleic Acids Res* **45**: D353–D361.

Konstantinidis, K.T., and Tiedje, J.M. (2004) Trends between gene content and genome size in prokaryotic species with

larger genomes. *Proc Natl Acad Sci U S A* **101**: 3160–3165.

Laliberté, E., and Legendre, P. (2010) A distance-based framework for measuring functional diversity from multiple traits. *Ecology* **91**: 299–305.

Langille, M.G.I., Zaneveld, J., Caporaso, J.G., McDonald, D., Knights, D., Reyes, J.A., *et al*. (2013) Predictive functional profiling of microbial communities using 16S rRNA marker gene sequences. *Nat Biotechnol* **31**: 814–821.

Liaw, A., and Wiener, M. (2002) Classification and regression by randomForest. *R News* **2**: 18–22.

Lynch, M.D.J., and Neufeld, J.D. (2015) Ecology and exploration of the rare biosphere. *Nat Rev Microbiol* **13**: 217–229.

Malmstrom, R.R., Cottrell, M.T., Elifantz, H., and Kirchman, D.L. (2005) Biomass production and assimilation of dissolved organic matter by SAR11 bacteria in the Northwest Atlantic Ocean. *Appl Environ Microbiol* **71**: 2979–2986.

Masák, J., Čejková, A., Schreiberová, O., and Řezanka, T. (2014) Pseudomonas biofilms: possibilities of their control. *FEMS Microbiol Ecol* **89**: 1–14.

Molenaar, D., van Berlo, R., de Ridder, D., and Teusink, B. (2009) Shifts in growth strategies reflect tradeoffs in cellular economics. *Mol Syst Biol* **5**: 323.

Nemergut, D.R., Costello, E.K., Hamady, M., Lozupone, C., Jiang, L., Schmidt, S.K., *et al*. (2011) Global patterns in the biogeography of bacterial taxa. *Environ Microbiol* **13**: 135–144.

O'Leary, N.A., Wright, M.W., Brister, J.R., Ciufo, S., Haddad, D., McVeigh, R., *et al*. (2015) Reference sequence (RefSeq) database at NCBI: current status, taxonomic expansion, and functional annotation. *Nucleic Acids Res* **44**: D733–D745.

Pedrós-Alió, C. (2012) The rare bacterial biosphere. *Ann Rev Mar Sci* **4**: 449–466.

Peter, H., Beier, S., Bertilsson, S., Lindström, E.S., Langenheder, S., and Tranvik, L.J. (2011) Function-specific response to depletion of microbial diversity. *ISME J* **5**: 351–361.

Ravintheran, S.K., Sivaprakasam, S., Loke, S., Lee, S.Y., Manickam, R., Yahya, A., *et al*. (2019) Complete genome sequence of Sphingomonas paucimobilis AIMST S2, a xenobiotic-degrading bacterium. *Sci Data* **6**: 280.

Rivett, D.W., and Bell, T. (2018) Abundance determines the functional role of bacterial phylotypes in complex communities. *Nat Microbiol* **3**: 767–772.

Roller, B.R.K., Stoddard, S.F., and Schmidt, T.M. (2016) Exploiting rRNA operon copy number to investigate bacterial reproductive strategies. *Nat Microbiol* **1**: 16160.

Schimel, J., and Schaeffer, N. (2012) Microbial control over carbon cycling in soil. *Front Microbiol* **3**.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., *et al*. (2003) Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res* **13**: 2498–2504.

Smriga, S., Fernandez, V.I., Mitchell, J.G., and Stocker, R. (2016) Chemotaxis toward phytoplankton drives organic matter partitioning among marine bacteria. *Proc Natl Acad Sci U S A* **113**: 1576–1581.

Steen, A.D., Crits-Christoph, A., Carini, P., DeAngelis, K.M., Fierer, N., Lloyd, K.G., and Cameron Thrash, J. (2019) High proportions of bacteria and archaea across most biomes remain uncultured. *ISME J* **13**: 3126–3130.

Vandera, E., Samiotaki, M., Parapouli, M., Panayotou, G., and Koukkou, A.I. (2015) Comparative proteomic analysis of Arthrobacter phenanthrenivorans Sphe3 on phenanthrene, phthalate and glucose. *J Proteomics* **113**: 73–89.

Vitousek, P.M., and Hooper, D.U. (1994) Biological diversity and terrestrial ecosystem biogeochemistry. In Schulze, E.-D., and Mooney, H.A. (eds). *Biodiversity and Ecosystem Function* Springer-Verlag, Berlin, Heidelberg, pp. 3–14.

Xiong, C., He, J.-Z., Singh, B.K., Zhu, Y.-G., Wang, J.-T., Li, P.-P., *et al*. (2021) Rare taxa maintain the stability of crop mycobiomes and ecosystem functions. *Environ Microbiol* **23**: 1907–1924.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F.O., Ludwig, W., Schleifer, K.-H., *et al*. (2014) Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nat Rev Microbiol* **12**: 635–645.

Supporting Information

Additional Supporting Information may be found in the online version of this article at the publisher's web-site:

Appendix S1: Supporting Information.