

Lawrence Berkeley National Laboratory

Recent Work

Title

Global diversity and biogeography of bacterial communities in wastewater treatment plants.

Permalink

<https://escholarship.org/uc/item/25g7w09b>

Journal

Nature microbiology, 4(7)

ISSN

2058-5276

Authors

Wu, Linwei
Ning, Daliang
Zhang, Bing
[et al.](#)

Publication Date

2019-07-01

DOI

10.1038/s41564-019-0426-5

Supplemental Material

<https://escholarship.org/uc/item/25g7w09b#supplemental>

Peer reviewed

Global Diversity and Biogeography of Bacterial Communities in Wastewater Treatment Plants

Linwei Wu^{1,2†}, Daliang Ning^{2,3,1†}, Bing Zhang^{1,2†}, Yong Li⁴, Ping Zhang^{2,5}, Xiaoyu Shan¹, Qiuting Zhang¹, Mathew Brown⁶, Zhenxin Li⁷, Joy D. Van Nostrand², Fangqiong Ling⁸, Naijia Xiao^{2,3}, Ya Zhang², Julia Vierheilig^{9,10}, George F. Wells¹¹, Yunfeng Yang¹, Ye Deng^{12,13}, Qichao Tu¹², Aijie Wang¹³, Global Water Microbiome Consortium[‡], Tong Zhang¹⁴, Zhili He^{15,16}, Jurg Keller¹⁷, Per H. Nielsen¹⁸, Pedro J. J. Alvarez¹⁹, Craig S. Criddle²⁰, Michael Wagner⁹, James M. Tiedje²¹, Qiang He^{22,23*}, Thomas P. Curtis^{6*}, David A. Stahl²⁴, Lisa Alvarez-Cohen^{25,26}, Bruce E. Rittmann²⁷, Xianghua Wen^{1*}, and Jizhong Zhou^{2,1,26*}

¹State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China; ²Institute for Environmental Genomics, Department of Microbiology and Plant Biology, and School of Civil Engineering and Environmental Sciences, University of Oklahoma, Norman, OK, USA; ³Consolidated Core Laboratory, University of Oklahoma, Norman, Oklahoma, USA; ⁴College of Resource & Environment Southwest University, Chongqing, China; ⁵Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA; ⁶School of Engineering, Newcastle University, Newcastle upon Tyne, UK; ⁷School of Environment, Northeastern Normal University, Changchun, China; ⁸Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, MO, USA; ⁹Department of Microbiology and Ecosystem Science, Division of Microbial Ecology, Research Network 'Chemistry meets Microbiology', University of Vienna, Vienna, Austria; ¹⁰Karl Landsteiner University of Health Sciences, Division of Water Quality and Health, Krems, Austria & Interuniversity Cooperation Centre for Water and Health; ¹¹Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL, USA; ¹²Institute for Marine Science and Technology, Shandong University, Qingdao, China; ¹³Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China; ¹⁴Environmental Biotechnology Laboratory, The University of Hong Kong, Hong Kong, China; ¹⁵Environmental Microbiomics Research Center, School of Environmental Science and Engineering, Sun Yat-Sen University, Guangzhou, China; ¹⁶Guangdong Provincial Key Laboratory of Environmental Pollution Control and Remediation Technology, Guangzhou, China; ¹⁷Advanced Water Management Centre, The University of Queensland, Brisbane, QLD, Australia; ¹⁸Department of Chemistry and Bioscience, Center for Microbial Communities, Aalborg University, Aalborg, Denmark; ¹⁹Department of Civil and Environmental Engineering, Rice University, Houston, TX, USA; ²⁰Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA; ²¹Center for Microbial Ecology, Michigan State University, East Lansing, MI, USA; ²²Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN, USA; ²³Institute for a Secure and Sustainable Environment, The University of Tennessee, Knoxville, TN, USA; ²⁴Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA; ²⁵Department of Civil and Environmental Engineering, College of Engineering, University of California, Berkeley, CA, USA; ²⁶Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; ²⁷Biodesign Swette Center for Environmental Biotechnology, Arizona State University, Tempe, AZ, USA.

† These authors contributed equally to this work.

‡ Other Global Water Microbiome Consortium members are listed at the end of this paper.

* Corresponding authors. To whom correspondence should be addressed regarding analysis, synthesis, and reprints, E-mail: jzhou@ou.edu. To whom correspondence should be addressed regarding experimental design and sampling, E-mails: xhwen@tsinghua.edu.cn, tom.curtis@newcastle.ac.uk, qianghe@utk.edu.

Microorganisms in wastewater treatment plants (WWTPs) are essential for water purification to protect public and environmental health. However, their diversity and the factors that control it are poorly understood. Using a systematic global-sampling effort, we analyzed the 16S rRNA gene sequences from ~1,200 activated sludge samples taken from 269 WWTPs in 23 countries on 6 continents. Our analyses revealed that the global activated sludge bacterial communities contain ~1 billion bacterial phylotypes with a Poisson lognormal diversity distribution. Despite this high diversity, activated sludge has a small global core bacterial community (n = 28 OTUs) that is strongly linked to activated sludge performance. Meta-analyses with global datasets associate the activated sludge microbiomes most closely to freshwater populations. In contrast to macroorganism diversity, activated sludge bacterial communities show no latitudinal gradient. Furthermore, their spatial turnover is scale-dependent and appears to be largely driven by stochastic processes (dispersal, drift), although deterministic factors (temperature, organic input) also are important. Our findings enhance mechanistic understanding of the global diversity and biogeography of activated sludge bacterial communities within a theoretical ecology framework and have important implications for microbial ecology and wastewater treatment processes.

Introduction

Microorganisms, the most diverse group of life on Earth¹, play crucial roles in the biogeochemical cycling of carbon (C), nitrogen (N), sulfur (S), phosphorus (P), and various metals. Unraveling the mechanisms generating and underlying microbial biodiversity is key to predicting ecosystem responses to environmental changes² and improving bioprocesses, such as wastewater treatment and soil remediation³. With recent advances in metagenomic technologies⁴, microbial biodiversity and distribution are being intensively studied in a wide variety of environments⁵⁻⁷, including the human gut, oceans, freshwater, air, and soil. However, we are just beginning to understand the diversity and biogeography of microbial communities in wastewater treatment plants (WWTPs)^{3,8}.

More than 300 km³ of wastewater is produced globally each year⁹. This volume equals one seventh of the global river volume¹⁰. About 60% of this wastewater is treated prior to release, and biological processes such as activated sludge are widely used in WWTPs⁹. Activated sludge employs microbial flocs or granules to remove C, N, P, micropollutants (e.g., toxins, pesticides, hormones, pharmaceuticals), and pathogens¹¹. Activated sludge relies on complex and incompletely defined microbial communities. As the largest application of biotechnology in the world¹², activated sludge is a vital infrastructure of modern urban societies¹³. Despite recent advances in understanding the microbial ecology of activated sludge¹⁴⁻¹⁶, the global picture of microbial diversity and distribution remains elusive. This information is essential to resolving controversies concerning the relative importance of stochastic versus deterministic community assembly in activated sludge³. Such information is also important for identifying key players in the process and providing a basis for targeted manipulation of activated sludge microbiomes.

We created a Global Water Microbiome Consortium (GWMC) (<http://gwmc.ou.edu/>) and conducted a global campaign for systematically collecting and analyzing activated sludge microbiomes. We collected activated sludge samples from 269 WWTPs in 86 cities, 23 countries, and 6 continents (Fig. 1a, Supplementary Table 1). Deep sequencing and analysis of 16S rRNA genes were performed to address fundamental ecological questions, including: (i) What is the extent of global diversity of activated sludge microbial communities? (ii) Does a core microbiome exist in activated sludge processes across different continents? (iii) Do activated sludge microbiomes show a latitudinal diversity gradient (LDG)? (iv) Is microbial biodiversity important to function in activated sludge processes? and (v) What is the relative importance of deterministic versus stochastic factors in regulating the composition, distribution, and functions of activated sludge microbial communities?

Species abundance distributions

Species abundance distribution (SAD), a universal tool in ecology¹⁷ and central to biodiversity theory, has not been rigorously tested in microbial ecology until recently¹⁸. Here, we tested common SAD models, including Poisson lognormal, log-series, Broken-stick, and Zipf. The Poisson lognormal model explained 99% of the variation of the activated sludge bacterial SADs, compared with 72% for log-series, 94% for Zipf model, and 14% for Broken-stick (Fig. 1b; Supplementary Table 2). Consistent with previous studies¹⁸, the Poisson lognormal model gave the best fit to the observed SADs.

Extent of global microbial diversity

One grand challenge in biodiversity research is determining the number of species in an ecological system¹⁹. We estimated the global richness of activated sludge bacterial communities based on two parameters^{19,20}. One is the total number of individuals (N_T), which was estimated as $4 - 6 \times 10^{23}$ bacteria in the global activated sludge community, based on published data⁹. The other is the quantity of the most abundant taxa (N_{\max}), which can be estimated based on either our sequence data or the dominance-scaling law¹⁹. The lognormal model predicts $1.1 (\pm 0.07) \times 10^9$ species in activated sludge systems globally, with N_{\max} at 1.2% of N_T based on our sequence data. The number of species increases only slightly, to $2.0 (\pm 0.2) \times 10^9$ species, using $N_{\max} = 0.4 \times N_T^{0.93}$ from the dominance-scaling law¹⁹ (Fig. 1c). The estimates of global activated sludge bacterial richness are only about one order of magnitude lower than that of the global ocean microbiome¹⁹ ($\sim 10^{10}$), even though the world's oceans represent an enormously larger ecosystem, which could be attributed to the higher volumetric productivity, thus higher concentration of bacterial cells, in activated sludge.

Global core bacterial community

Previous studies have reported the core community in WWTPs at regional scales. For example, core genera existed in Danish¹⁴ and Asian¹⁵ WWTPs, but less than 10% of the genera overlapped. Thus, a global core cannot be established from those regional studies.

At the global scale, occupancy-frequency and occupancy-abundance analyses revealed a hyper-dominant pattern (Supplementary Fig. 1a) in which the 866 most abundant OTUs (1.39% of the total OTU number) accounted for 50.06% of the total abundance. Similar hyper-dominance patterns were observed in other macro-²¹ and microbial communities²².

A core bacterial community was determined based on abundance and occurrence frequency of OTUs (see Methods for details). About 0.05% (28 OTUs) constituted a global core that accounted for $12.4\% \pm 0.2\%$ (mean \pm SE) of the sequences in activated sludge samples (Fig. 2a; Supplementary Table 3). Most (82%) of the core community members belonged to *Proteobacteria*, with 15 OTUs classified as β -*Proteobacteria* (Fig. 2b). The most abundant OTU, accounting for $1.14\% \pm 0.05\%$ of the sequence abundance in activated sludge samples and occurring in 85% of all samples, was 99% similar to the γ -proteobacterium *Dokdonella kunshanensis* DC-3²³. The second most abundant OTU ($0.89\% \pm 0.06\%$ in relative abundance and occurring in 96% of all samples) belonged to *Zoogloea*, a dominant genus in activated sludge communities¹⁵, with *Z. ramigera* known to enhance the flocculation of activated sludge²⁴. A *Nitrospira* OTU (OTU_6) was also identified as a core taxon, reflecting its importance for nitrite oxidation or complete ammonia oxidation in activated sludge^{25,26}. OTU_7 is closely related to *Arcobacter* species, which are highly abundant in raw sewage²⁷ and include potential pathogens, such as *A. cryaerophilus*, *A. butzleri*, and *A. skirrowii*²⁸. Furthermore, two putative polyphosphate- accumulating organisms (PAOs), a “*Candidatus* Accumulimonas” OTU (OTU_37) and a “*Candidatus* Accumulibacter” OTU (OTU_25), were identified as core taxa, although only 149 out of the 269 sampled WWTPs operate as enhanced biological P removal

(EBPR) systems. Apparently, “*Candidatus Accumulimonas*” and “*Candidatus Accumulibacter*” exhibit some metabolic versatility.

The global core community has some overlap with previous studies. For example, *Zoogloea* species were proposed as core denitrifiers, and certain *Saprospiraceae* species play an important role in hydrolysis in EBPR systems²⁹. However, some discrepancies also occurred. Saunders et al. showed *Nitrotoga* rather than *Nitrospira* as primary nitrite-oxidizers in Danish WWTPs¹⁴. Lawson et al. found low abundances of both *Nitrotoga* and *Nitrospira* in a pilot-scale EBPR treatment plant, but *Nitrotoga* maintained high potential activities based on high SSU rRNA:rDNA ratios³⁰. Regarding PAOs, we identified “*Candidatus Accumulimonas*” and “*Candidatus Accumulibacter*” as global core taxa, while *Tetrasphaera* was the core PAO in Danish WWTPs^{14,31}.

We similarly determined core communities for a variety of ecosystems at the global scale based on the Earth Microbiome Project (EMP) datasets⁵. Soil, human feces, air, and freshwater microbiomes had 9, 6, 2, and 1 bacterial OTUs identified as core taxa, respectively (Supplementary Table 4). No core taxa were found for animal feces and the ocean, possibly due to highly variable community compositions. Notably, the core community for activated sludge had no overlap with the other habitats, suggesting that activated sludge selects for a unique core community.

Latitudinal diversity pattern

Latitudinal diversity gradients (LDG), whereby species richness tends to decrease as latitude increases³², are well documented in plant and animal ecology³³. Recently, several studies examined LDG patterns in natural microbial communities, but found no clear trends^{6,7,34}. In contrast, activated sludge operates under relatively stable and similar conditions everywhere. Thus, one might not expect activated sludge microbial communities to exhibit LDG.

We examined the relationship between OTU richness and latitude. OTU richness peaked at intermediate latitude, with a mean air temperature $\sim 15^{\circ}\text{C}$ (Fig. 1d). As taxonomic and phylogenetic diversity were highly correlated ($R^2 = 0.92$), the trend was similar for phylogenetic diversity (Supplementary Fig. 2a). These results suggest that a LDG does not occur in activated sludge microbiomes; this parallels the global ocean microbiome⁷, but contrasts with some ocean³⁴ and soil communities³⁵. In addition, the relationship between bacterial richness and temperature (Supplementary Fig. 2b, c) did not fit predictions from the metabolic theory of ecology³⁶. This theory cannot explain bacterial richness based on air temperature (Supplementary Fig. 2b, $R^2 < 0.001$) and mixed liquid temperature (Supplementary Fig. 2c, $R^2 = 0.03$).

Continental-level differences in bacterial community structure

Variations in community composition (β -diversity) are key for understanding community assembly mechanisms^{2,37} and ecosystem functioning³⁸. To understand how the bacterial community composition of activated sludge varied across different spatial scales, we examined taxonomic and phylogenetic diversity. First, diversity was highest in Asia and lowest in South

America (Supplementary Table 5). Second, considerable variations between activated sludge samples were observed even at the phylum level (Supplementary Fig. 1b). Although the taxonomic and phylogenetic community structures were not clearly separated at the OTU level in two-dimensional ordinations (Supplementary Fig. 1c, d), PERMANOVA indicated that taxonomic and phylogenetic composition were significantly different ($P < 0.001$) between any two continents (Supplementary Table 6). Third, climate and activated sludge process type exerted significant effects ($P = 0.001$) on microbial community structure, but these were overwhelmed by continental geographical separation (Supplementary Table 7). For example, bacterial communities of the same climate type in North America and Asia were distinguished by their continental origins rather than being clustered together (Supplementary Fig. 1e, f). While the activated sludge bacterial communities had higher similarity to those of freshwater and soil than to other environments (Fig. 3a), they harbored a unique microbiome distinctly different from all other habitats (Supplementary Table 8).

A Bayesian approach³⁹ was employed to identify potential sources of activated sludge bacterial communities at the genus level. The most dominant potential source was freshwater, attributing on average 46% of genera, followed by soil (17% on average) and ocean (12% on average) (Fig. 3b). Apparently, environmental characteristics are more similar between an activated sludge bioreactor and freshwater than the others. Activated sludge and freshwater have potentially high immigration events through connected water systems, such as wastewater discharge to rivers after treatment.

Scale-dependent distance-decay patterns

Another fundamental pattern in ecology is the distance-decay relationship (DDR)^{17,40}, in which community similarity decreases as geographic distance increases. Consistent with results in other domains³⁷, we hypothesized that (i) the slope of the DDR curve would vary over local, regional, and global scales, and (ii) the spatial turnover rates of activated sludge microbial communities would be lower than those observed in natural habitats, especially for non-flowing ecosystems, such as soils⁴¹.

Supporting our first hypothesis, significant negative DDRs ($P < 0.001$) were observed across all scales based on taxonomic diversity (slope = -0.06 for Sorensen, -0.08 for Bray-Curtis, and -0.08 for Canberra distance) and phylogenetic diversity (slope = -0.04 for unweighted Unifrac, and -0.02 for weighted Unifrac) (Fig. 4a, Supplementary Table 9). The slopes of DDRs depended significantly on spatial scale. The DDR slopes across cities within a continent (-0.13 ~ -0.16 for taxonomic similarity indices; -0.03 ~ -0.09 for phylogenetic similarity indices) were significantly ($P = 0.001$) steeper (> 2 times) than the overall slopes for all similarity metrics (Supplementary Table 9). Countering our second hypothesis, the overall spatial turnover rates of the activated sludge communities were similar to those found in non-flowing natural habitats such as soils⁶ and sediments³⁷.

Relationships between the community structure and activated sludge functions

Understanding the relationships between biodiversity and ecosystem function is a critical topic in ecology⁴². Despite decades of intensive studies, the biodiversity-function relationship is still hotly debated, particularly in microbial ecology⁴³. A recent meta-analysis of the microbial ecology literature found that less than one-half of all mechanistic claims were backed up by any statistical tests⁴⁴. Since activated sludge is an engineered system, we hypothesized that there would be a strong linkage between the activated sludge bacterial community structure and its functions.

To assess functions, we calculated the removal rates of organic matter (biochemical oxygen demand (BOD), chemical oxygen demand (COD)), total phosphorus, total nitrogen, and ammonium nitrogen. Partial Mantel tests revealed that the distance-corrected changes of activated sludge-community composition were significantly correlated with all measured removal rates ($P < 0.032$), except for the ammonium-nitrogen removal rate ($P > 0.18$) (Supplementary Table 10). Of the 28 global core OTUs, 27 were significantly correlated (adjusted $P < 0.05$) with at least one of the five functions examined. Most of the correlations (81%) were positive (Fig. 2c). Also, about 80% of the non-core OTUs showed significant correlations (adjusted $P < 0.05$) with at least one function, and 40% of these correlations were positive (Supplementary Fig. 3a). All of these results indicated that the structure of the activated sludge bacterial communities, particularly the dominant populations, is critical to maintaining activated sludge functions.

The global dataset also allows us to assess the importance of specific functional groups to activated sludge functions. The nitrifying microbial community, including *Nitrospira* and

Nitrosomonas OTUs, showed a closer correlation with the ammonium- nitrogen removal rate than did the whole community (Supplementary Table 10; P of Bray-Curtis distance =0.04). Further analysis revealed significant positive correlations of *Nitrospira* (Spearman's $\rho = 0.40$, adjusted $P < 0.001$) and *Nitrosomonas* (Spearman's $\rho = 0.21$, adjusted $P < 0.001$) abundance to the percentages of ammonium-nitrogen removal (% of influent concentration), but not to the ammonium-nitrogen removal rate (Supplementary Fig. 3b). *Nitrospira* was the top genus correlating with the percentage of ammonium-nitrogen removal, corroborating its role in nitrite oxidation in activated sludge. Regarding ammonium-oxidizing bacteria (AOB), an activated sludge bioreactor harboured 15 *Nitrosomonas* OTUs on average, which made up $0.73\% \pm 0.06\%$ of the sequence abundance (Supplementary Table 11).

Consistent with our expectation, the activated sludge community composition was significantly correlated with the TP removal rate for the samples from EBPR plants, but not for non-EBPR plants (Supplementary Table 10), as P removal processes in non-EBPR plants are predominantly chemical. The diversity of the three potential PAOs³¹ were significantly different ($P < 0.0001$, two tailed paired-t test between any two organisms): 8.2 ± 0.2 “*Candidatus Accumulimonas*” OTUs, 6.6 ± 0.2 “*Candidatus Accumulibacter*” OTUs, and 3.2 ± 0.1 *Tetrasphaera* OTUs within a typical activated sludge bioreactor. While the relative abundance of “*Candidatus Accumulimonas*” ($0.42\% \pm 0.06\%$) was not different from that of “*Candidatus Accumulibacter*” ($0.42\% \pm 0.04\%$) (two tailed paired-t test, $P = 0.92$), both were more abundant than *Tetrasphaera* (mean relative abundance $0.17\% \pm 0.02\%$) (two tailed paired-t test, $P < 0.0001$) (Supplementary Table 12).

299

300 **Stochastic community assembly**

301

302 Since WWTPs are well-controlled engineered ecosystems, we hypothesized that the activated
303 sludge community assembly has a deterministic nature, and we calculated the null model-based
304 stochastic ratios⁴¹ with taxonomic and phylogenetic metrics. The average stochastic ratios based
305 on these four metrics all were higher than 0.75 (Fig. 4b), suggesting that stochastic factors were
306 more important than deterministic factors in influencing community composition, at least
307 partially contradicting our hypothesis.

308

309 To discern the relative importance of various factors contributing to spatial turnover of the
310 activated sludge bacterial communities, we performed multiple ‘regression on matrices’ (MRM)
311 analyses and a subsequent variance partition analysis (VPA) based on various taxonomic and
312 phylogenetic diversity metrics (Fig. 4c, Supplementary Fig. 4). Over all scales, the MRM model
313 explained considerable and significant portions of the community variations based on Bray-
314 Curtis similarity ($R^2 = 0.46$, $P = 0.001$) (Fig. 4c), with >50% variations unexplained. Among
315 these, 25%, 11%, and 10% of the variations were explained by geographical distance,
316 environmental variables, and their interactions, respectively (Fig. 4c). Similar trends were
317 observed across different scales, with environmental variables explaining < 30% of community
318 variations based on different similarity metrics (Supplementary Fig. 4). These results support
319 those inferred from the null-model-based stochastic ratio analysis.

320

Environmental drivers of community composition

Because both stochastic and deterministic factors are important in forming the activated sludge community assembly, we attempted to discern the roles of individual deterministic factors in shaping community structure. We correlated the geographic distance-corrected dissimilarities of community composition with those of environmental variables by the partial Mantel test (Supplementary Fig. 5a, Supplementary Table 13). Overall, the microbial community composition had strong correlations with absolute latitude, mean annual temperature (MAT), solids retention time (SRT, the average time which activated sludge solids are in the system), and influent COD and BOD concentrations, representing organic matter ($r_m = 0.23-0.30$, $P = 0.001$).

More in-depth analysis by structural equation modeling (SEM) revealed direct and indirect effects of the environmental drivers (Fig. 5a). Consistent with the Mantel test, temperature had the strongest direct effects on PC1 representing the community structure (standardized path coefficient, $\beta = 0.50$, $P < 0.001$). It also had weak negative impacts on species richness ($\beta = -0.14$, $P < 0.001$). This is consistent with previous observations at local^{45,46} and regional⁴⁷ scales that highlighted temperature as a key factor influencing activated sludge community structure and, in particular, abundance and diversity of slow-growing microorganisms such as AOB and nitrite oxidizing bacteria (NOB).

Various biotic and abiotic factors (e.g., food-to-microorganisms ratio [F/M] (the ratio of organic matter to microorganisms), dissolved oxygen concentration, and SRT) directly affected BOD-removal rates (Fig. 5a). Influent BOD likely has an impact on bacterial composition through its

effect on the F/M ratio ($\beta = 0.31$, $P < 0.001$), which is inversely related to the SRT. Influent BOD is the most influential environmental variable directly related to bacterial richness ($\beta = -0.28$, $P < 0.001$), and the abundance-weighted mean rRNA gene copy number significantly increased with the influent BOD ($R^2 = 0.19$, $P < 0.0001$; Fig. 5b). All of these results are consistent with the resource-competition theory⁴⁸, which predicts that high species diversity occurs with low to intermediate supply of resources, but fast-growing r-strategists outcompete efficient-scavenging K-strategists at high resource levels⁴⁹.

To independently test the strength of correlation for each of the three strongest parameters (temperature, SRT, and influent BOD) with bacterial community structure, we performed random-forest analysis, a machine learning-based method. Using species abundance as the input data, the model predicted temperature, SRT, and influent BOD with an explained variance of 69%, 25%, and 18%, respectively (Fig. 5c, Supplementary Fig. 5b). When controlling for spatial auto-correlation, models of temperature continued to have higher accuracy (Supplementary Fig. 5b). For example, the America-fitted model of temperature, i.e., a model trained solely by North and South America samples, was able to capture variations in the temperatures of Asia samples (cross-validated $R^2 = 0.47$) (Fig. 5c). The random-forest model also revealed the most important OTUs for predicting temperature (Supplementary Fig. 5c). These results corroborate that temperature is the major environmental variable shaping the activated sludge bacterial compositions at the global scale, although it only has a weak effect on species richness (Fig. 5a).

Conclusions and future perspectives

Through well-coordinated international efforts, we systematically examined global diversity and biogeography of activated sludge bacterial communities within the context of theoretical ecology frameworks. Our findings enhance understanding of microbial ecology in activated sludge, setting the stage for various future analyses of WWTP microbiomes, as well as other microbial communities that span the globe.

Based on experimental and theoretical analyses, we estimate that activated sludge systems are globally inhabited by $\sim 10^9$ different bacterial species. In contrast, only about 10^4 species have been cultivated and studied in detail¹⁹. If we assume that all cultivated species are present in activated sludge, potentially 99.999% of activated sludge microbial taxa remain uncultured. Although more and more microorganisms have been genomically characterized, exploring physiological attributes, which requires cultivation, represents a formidable task for future microbiologists and process engineers⁵⁰. This finding also highlights how little we know of the world's microbiome, even in one of the most common and well-controlled systems in the built environment. Despite the very large diversity in activated sludge, a functionally important global core community consists of fewer than 30 taxa. This core might serve as the “most wanted” list for future experimental efforts to understand their genetic, biochemical, physiological, and ecological traits.

Even though activated sludge is a managed ecosystem, its bacterial composition appears to be driven most likely by stochastic processes, such as dispersal and drift, which apparently contradicts conventional wisdom. However, deterministic factors (e.g., temperature, SRT, and organic C inputs) play important roles in regulating the structure of the activated sludge

community. This could be important for developing operating strategies to maintain biodiversity that promotes stable system performance. Perhaps one could overcome dispersal limitation by establishing WWTPs, or repopulating failed WWTPs using an inoculum of activated sludge from functioning WWTPs, which is a common practice in environmental engineering. Alternately, one could alternate organic C loadings and/or operational conditions to manipulate the activated sludge community's structure to select for the microorganisms having the desired functions.

Finally, apart from the practical implications of this study, it appears that the global bacterial communities in activated sludge follow various macroecological patterns, such as SADs, DDRs, resource theory, and community assembly mechanisms. Given that activated sludge can be controlled and monitored, it could be an excellent system for testing how well different macroecological theories apply to microbial ecology: e.g., the relationships among biodiversity, food-web interactions, succession, stability, and ecosystem functioning.

References

- 1 Torsvik, V., Øvreås, L. & Thingstad, T. F. Prokaryotic diversity--magnitude, dynamics, and controlling factors. *Science* **296**, 1064-1066 (2002).
- 2 Chase, J. M. & Myers, J. A. Disentangling the importance of ecological niches from stochastic processes across scales. *Philos Trans R Soc Lond B Biol Sci* **366**, 2351-2363 (2011).
- 3 Ofiteru, I. D. *et al.* Combined niche and neutral effects in a microbial wastewater treatment community. *Proc Natl Acad Sci USA* **107**, 15345-15350 (2010).
- 4 Zhou, J. *et al.* High-Throughput Metagenomic Technologies for Complex Microbial Community Analysis: Open and Closed Formats. *mBio* **6**, e02288-02214 (2015).
- 5 Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457-463, doi:10.1038/nature24621 (2017).
- 6 Fierer, N. & Jackson, R. B. The diversity and biogeography of soil bacterial communities. *Proc Natl Acad Sci USA* **103**, 626-631 (2006).
- 7 Sunagawa, S. *et al.* Structure and function of the global ocean microbiome. *Science* **348**, 1261359 (2015).
- 8 National Academies of Sciences, E. & Medicine. *Microbiomes of the built environment: a research agenda for indoor microbiology, human health, and buildings*. (National Academies Press, 2017).
- 9 Mateo-Sagasta, J., Raschid-Sally, L. & Thebo, A. in *Wastewater* 15-38 (Springer, 2015).
- 10 Gleick, P. H. Water resources. *Encyclopedia of climate and weather* **2**, 817-823 (1996).
- 11 van Loosdrecht, M. C. & Brdjanovic, D. Anticipating the next century of wastewater treatment. *Science* **344**, 1452-1453 (2014).
- 12 Xia, S. *et al.* Bacterial community structure in geographically distributed biological wastewater treatment reactors. *Environ Sci Technol* **44**, 7391-7396 (2010).
- 13 Grant, S. B. *et al.* Taking the “waste” out of “wastewater” for human water security and ecosystem sustainability. *Science* **337**, 681-686 (2012).
- 14 Saunders, A. M., Albertsen, M., Vollertsen, J. & Nielsen, P. H. The activated sludge ecosystem contains a core community of abundant organisms. *ISME J* **10**, 11 (2016).
- 15 Zhang, T., Shao, M.-F. & Ye, L. 454 Pyrosequencing reveals bacterial diversity of activated sludge from 14 sewage treatment plants. *ISME J* **6**, 1137-1147 (2012).
- 16 Wagner, M. & Loy, A. Bacterial community composition and function in sewage treatment systems. *Curr Opin Biotechnol* **13**, 218-227 (2002).
- 17 Morlon, H. *et al.* Spatial patterns of phylogenetic diversity. *Ecol Lett* **14**, 141-149 (2011).
- 18 Shoemaker, W. R., Locey, K. J. & Lennon, J. T. A macroecological theory of microbial biodiversity. *Nat Ecol Evol* **1**, 107, doi:10.1038/s41559-017-0107 (2017).
- 19 Locey, K. J. & Lennon, J. T. Scaling laws predict global microbial diversity. *Proc Natl Acad Sci USA* **113**, 5970-5975 (2016).
- 20 Curtis, T. P., Sloan, W. T. & Scannell, J. W. Estimating prokaryotic diversity and its limits. *Proc Natl Acad Sci USA* **99**, 10494-10499 (2002).
- 21 Ter Steege, H. *et al.* Hyperdominance in the Amazonian tree flora. *Science* **342**, 1243092 (2013).
- 22 De Vargas, C. *et al.* Eukaryotic plankton diversity in the sunlit ocean. *Science* **348**, 1261605 (2015).

448 23 Li, Y. *et al.* Dokdonella kunshanensis sp. nov., isolated from activated sludge, and
449 emended description of the genus Dokdonella. *Int J Syst Evol Microbiol* **63**, 1519-1523
450 (2013).

451 24 Rosselló-Mora, R. A., Wagner, M., Amann, R. & Schleifer, K.-H. The abundance of
452 Zoogloea ramigera in sewage treatment plants. *Appl Environ Microbiol* **61**, 702-707
453 (1995).

454 25 Daims, H. *et al.* Complete nitrification by Nitrospira bacteria. *Nature* **528**, 504 (2015).

455 26 Daims, H., Nielsen, J. L., Nielsen, P. H., Schleifer, K.-H. & Wagner, M. In Situ
456 Characterization of Nitrospira-Like Nitrite-Oxidizing Bacteria Active in Wastewater
457 Treatment Plants. *Appl Environ Microbiol* **67**, 5273-5284 (2001).

458 27 Fisher, J. C., Levican, A., Figueras, M. J. & McLellan, S. L. Population dynamics and
459 ecology of Arcobacter in sewage. *Front Microbiol* **5** (2014).

460 28 Collado, L. & Figueras, M. J. Taxonomy, epidemiology, and clinical relevance of the
461 genus Arcobacter. *Clin Microbiol Rev* **24**, 174-192 (2011).

462 29 Nielsen, P. H., Saunders, A. M., Hansen, A. A., Larsen, P. & Nielsen, J. L. Microbial
463 communities involved in enhanced biological phosphorus removal from wastewater—a
464 model system in environmental biotechnology. *Curr Opin Biotechnol* **23**, 452-459 (2012).

465 30 Lawson, C. E. *et al.* Rare taxa have potential to make metabolic contributions in
466 enhanced biological phosphorus removal ecosystems. *Environ Microbiol* **17**, 4979-4993
467 (2015).

468 31 Stokholm-Bjerregaard, M. *et al.* A critical assessment of the microorganisms proposed to
469 be important to enhanced biological phosphorus removal in full-scale wastewater
470 treatment systems. *Front Microbiol* **8**, 718 (2017).

471 32 Hillebrand, H. On the generality of the latitudinal diversity gradient. *Am Nat* **163**, 192-
472 211 (2004).

473 33 Martiny, J. B. H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat*
474 *Rev Microbiol* **4**, 102-112 (2006).

475 34 Fuhrman, J. A. *et al.* A latitudinal diversity gradient in planktonic marine bacteria. *Proc*
476 *Natl Acad Sci USA* **105**, 7774-7778, doi:10.1073/pnas.0803070105 (2008).

477 35 Zhou, J. *et al.* Temperature mediates continental-scale diversity of microbes in forest
478 soils. *Nat Commun* **7**, 12083, doi:10.1038/ncomms12083 (2016).

479 36 Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M. & West, G. B. Toward a
480 metabolic theory of ecology. *Ecology* **85**, 1771-1789 (2004).

481 37 Martiny, J. B., Eisen, J. A., Penn, K., Allison, S. D. & Horner-Devine, M. C. Drivers of
482 bacterial beta-diversity depend on spatial scale. *Proc Natl Acad Sci USA* **108**, 7850-7854,
483 doi:10.1073/pnas.1016308108 (2011).

484 38 Zhou, J. *et al.* Stochastic assembly leads to alternative communities with distinct
485 functions in a bioreactor microbial community. *mBio* **4**, e00584-00512 (2013).

486 39 Knights, D. *et al.* Bayesian community-wide culture-independent microbial source
487 tracking. *Nat Methods* **8**, 761-763 (2011).

488 40 Zhou, J. & Ning, D. Stochastic Community Assembly: Does It Matter in Microbial
489 Ecology? *Microbiol Mol Biol Rev* **81**, e00002-00017 (2017).

490 41 Zhou, J. *et al.* Stochasticity, succession, and environmental perturbations in a fluidic
491 ecosystem. *Proc Natl Acad Sci USA* **111**, E836-E845 (2014).

492 42 Hooper, D. U. *et al.* A global synthesis reveals biodiversity loss as a major driver of
493 ecosystem change. *Nature* **486**, 105 (2012).

- 43 Krause, S. *et al.* Trait-based approaches for understanding microbial biodiversity and ecosystem functioning. *Front Microbiol* **5**, 251 (2014).
- 44 Bier, R. L. *et al.* Linking microbial community structure and microbial processes: an empirical and conceptual overview. *FEMS Microbiol Ecol* **91**, doi:10.1093/femsec/fiv113 (2015).
- 45 Wells, G.F. *et al.* Ammonia-oxidizing communities in a highly aerated full-scale activated sludge bioreactor: betaproteobacterial dynamics and low relative abundance of Crenarchaea. *Environ Microbiol* **11**, 2310-2328 (2009).
- 46 Karkman, A., Mattila, K., Tamminen, M. & Virta, M. Cold temperature decreases bacterial species richness in nitrogen-removing bioreactors treating inorganic mine waters. *Biotechnol Bioeng* **108**, 2876-2883 (2011).
- 47 Griffin, J.S. & Wells, G.F. Regional synchrony in full-scale activated sludge bioreactors due to deterministic microbial community assembly. *ISME J* **11**, 500-511 (2017).
- 48 Tilman, D. *Resource competition and community structure*. (Princeton university press, 1982).
- 49 Wu, L. *et al.* Microbial functional trait of rRNA operon copy numbers increases with organic levels in anaerobic digesters. *ISME J* **11**, 2874-2878, doi:10.1038/ismej.2017.135 (2017).
- 50 Pedrós-Alió, C. & Manrubia, S. The vast unknown microbial biosphere. *Proc Natl Acad Sci USA* **113**, 6585-6587 (2016).

Materials & Correspondence

To whom correspondence should be addressed regarding analysis, synthesis and reprints, E-mail: jzhou@ou.edu. To whom correspondence should be addressed regarding experimental design and sampling, E-mails: xhwen@tsinghua.edu.cn, tom.curtis@newcastle.ac.uk, qianghe@utk.edu.

Acknowledgements

We thank Teresa Allen, Ahmed Al-Omari, Ryan Bart, David Crowley, George Harwood, Tom Hensley, Shiaw-Jy Huitric, Margarida M.L. Martins, Alex Mena, Bipin Pathak, Sofia Pereira, David E. Sauble, Mike Taylor, Phuong Truong, Dan VanderSchuur, Anabela Vieira, and Daniela Zambrano for helping with sampling and metadata collection. This work was supported by Tsinghua University Initiative Scientific Research Program (No. 20161080112), the National Scientific Foundation in China (51678335), the State Key Joint Laboratory of Environmental Simulation and Pollution Control (18L02ESPC) in China, and the Office of the Vice President for Research at the University of Oklahoma. LW.W. and B.Z. were generously supported by China Scholarship Council (CSC).

Author contributions

All authors contributed experimental assistance and intellectual input to this study. The original concept was conceived by JZ.Z.; Experimental strategies and sampling design were developed by JZ.Z., XH.W., T.P.C., Q.H., ZL.H., and DL.N.; Sample collections were coordinated by Q.H., DL.N., XH.W., T.P.C., B.Z., M.B., G.F.W., JZ.Z., and other GWMC members. J.D.V.N and DL.N. managed shipping. Yo.L., B.Z., ZX.L., DL.N., and some GWMC members (F.B., S.K., J.V., A.N.R., D.D.C.V., C.E., L.C., J.C.A., C.D.L., L.C.M., A.C., Pa.B., D.A.) did DNA extraction. P.Z. performed DNA sequencing with the help from LY.W.; LW.W., DL.N., JZ.Z., B.Z., XY.S., QT.Z., FQ.L., NJ.X., and RM.T. performed data analysis with help from Y.D., QC.T., T.Z., Ya.Z and AJ.W.; LW.W., JZ.Z., and DL.N. wrote the manuscript with the help from B.E.R., L.A.-C., M.W., C.S.C., D.A.S., G.F.W., J.M.T., P.J.J.A., J.K., J.V., P.H.N., R.G.L., XH.W., ZL.H., and YF.Y..

Global Water Microbiome Consortium

Members besides the authors listed on the first page:

Dany Acevedo¹, Miriam Agullo-Barcelo², Gary L. Andersen^{3,4}, Juliana Calabria de Araujo⁵, Kevin Boehnke⁶, Philip Bond², Charles B. Bott⁷, Patricia Bovio⁸, Rebecca K. Brewster⁶, Faizal Bux⁹, Angela Cabezas⁸, Léa Cabrol^{10,11}, Si Chen¹², Ting Chen^{13,14}, Claudia Etchebehere⁸, Amanda Ford⁷, Dominic Frigon¹⁵, Janeth Sanabria Gómez¹, James S. Griffin¹⁶, April Z. Gu¹⁷, Moshe Habagil¹⁸, Lauren Hale¹⁹, Steven D. Hardeman²⁰, Marc Harmon²¹, Harald Horn²², Zhiqiang Hu²³, Shameem Jauffur^{15,24}, David R. Johnson²⁵, Alexander Keucken^{18,26}, Sheena Kumari⁹, Cintia Dutra Leal⁵, Laura A. Lebrun²⁷, Jangho Lee²⁸, Minjoo Lee²⁸, Zarraz MP Lee²⁹, Mengyan Li³⁰, Xu Li³¹, Yu Liu^{32,29}, Richard G. Luthy³³, Leda C. Mendonça-Hagler³⁴, Francisca Gleire Rodriguez de Menezes³⁵, Arthur J. Meyers³⁶, Amin Mohebbi^{31,37}, Adrian Oehmen³⁸, Andrew Palmer³⁹, Prathap Parameswaran⁴⁰, Joonhong Park²⁸, Deborah Patsch²⁵, Valeria Reginato⁴¹, Francis L. de los Reyes III⁴², Adalberto Noyola Robles⁴³, Simona Rossetti⁴⁴,

560 Jatinder Sidhu³⁹, William T. Sloan⁴⁵, Kylie Smith³⁹, Oscarina Viana de Sousa³⁵, Kyle Stephens⁴⁶,
561 Renmao Tian¹⁹, Nicholas B. Tooker¹⁷, Daniel De los Cobos Vasconcelos⁴³, Steve Wakelin⁴⁷, Bei
562 Wang⁴⁸, Joseph E. Weaver⁴², Stephanie West²², Paul Wilmes²⁷, Sung-Geun Woo³³, Jer-Horng
563 Wu⁴⁹, Liyou Wu¹⁹, Chuanwu Xi⁶, Meiying Xu⁵⁰, Tao Yan⁵¹, Min Yang⁵², Michelle Young⁴⁰,
564 Haowei Yue⁵³, Qian Zhang⁵¹, Wen Zhang⁵⁴, Yu Zhang⁵², Hongde Zhou⁴⁸

565
566 **Steering Committee:** Jizhong Zhou^{19,53,55}, Xianghua Wen⁵³, Thomas P. Curtis⁵⁶, Qiang He^{12,57},
567 Zhili He^{58,59}; **Contact:** Jizhong Zhou (jzhou@ou.edu), Daliang Ning^{19,60,53} (ningdaliang@ou.edu)

568
569 ¹Environmental Microbiology and Biotechnology Laboratory, Engineering School of
570 Environmental & Natural Resources, Engineering Faculty, Universidad del Valle–Sede
571 Meléndez, Cali, Colombia; ²Advanced Water Management Centre, The University of
572 Queensland, Brisbane, QLD, Australia; ³Department of Environmental Science, Policy, and
573 Management, University of California, Berkeley, USA; ⁴Environmental Genomics and Systems
574 Biology Division, Lawrence Berkeley National Laboratory, Berkeley, USA; ⁵Universidade
575 Federal de Minas Gerais, Departamento de Engenharia Sanitária e Ambiental, Brazil;
576 ⁶Department of Environmental Health Sciences, The University of Michigan, Ann Arbor, MI,
577 USA; ⁷Hampton Roads Sanitation District (HRSD), Virginia Beach, VA, USA; ⁸Microbial
578 Ecology Laboratory, Microbial Biochemistry and Genomics Department, Biological Research
579 Institute “Clemente Estable”, Montevideo, Uruguay; ⁹Institute of Water and Wastewater
580 Technology, Durban University of Technology, South Africa; ¹⁰Aix-Marseille University CNRS
581 IRD, MIO UM110 Mediterranean Institute of Oceanography, Marseille, France; ¹¹Escuela de
582 Ingeniería Bioquímica, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile;
583 ¹²Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville,
584 TN, USA; ¹³Department of Computer Science, Tsinghua University, Beijing, China;
585 ¹⁴Bioinformatics Division, Tsinghua National Laboratory of Information Science and
586 Technology, Tsinghua University, Beijing, China; ¹⁵Microbial Community Engineering
587 Laboratory, Department of Civil Engineering and Applied Mechanics, McGill University,
588 Montreal, Canada; ¹⁶Department of Chemical and Biological Engineering, Northwestern
589 University, Evanston, IL, USA; ¹⁷Department of Civil and Environmental Engineering,
590 Northeastern University, Boston, MA, USA; ¹⁸Vatten & Miljö i Väst AB (VIVAB), Falkenberg,
591 Sweden; ¹⁹Institute for Environmental Genomics, Department of Microbiology and Plant
592 Biology, and School of Civil Engineering and Environmental Sciences, University of Oklahoma,
593 Norman, OK, USA; ²⁰Norman Water Reclamation Facility, Norman, OK, USA; ²¹Golden Heart
594 Utilities, Fairbanks, AK, USA; ²²Karlsruhe Institute of Technology, Engler-Bunte-Institut, Water
595 Chemistry and Water Technology, Germany; ²³Department of Civil and Environmental
596 Engineering, University of Missouri, Columbia, MO, USA; ²⁴Department of Building, Civil and
597 Environmental Engineering, Concordia University, Montreal, Canada; ²⁵Department of
598 Environmental Microbiology, Eawag, Dübendorf, Switzerland; ²⁶Water Resources Engineering,
599 Faculty of Engineering, Lund University, Lund, Sweden; ²⁷Luxembourg Centre for Systems
600 Biomedicine, University of Luxembourg, Belvaux, Luxembourg; ²⁸Department of Civil and
601 Environmental Engineering, Yonsei University, Seoul, South Korea; ²⁹Advanced Environmental
602 Biotechnology Centre, Nanyang Environment & Water Research Institute, Nanyang
603 Technological University, Singapore; ³⁰Department of Civil and Environmental Engineering,
604 Rice University, Houston, TX, USA; ³¹Department of Civil Engineering, University of
605 Nebraska-Lincoln, NE, USA; ³²School of Civil and Environmental Engineering, Nanyang

Technological University, Singapore; ³³Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA; ³⁴Plant Biotechnology Program, Federal University of Rio de Janeiro, UFRJ, Brazil; ³⁵Federal University of Ceará, UFC, Brazil; ³⁶University of Tennessee, Center for Environmental Biotechnology, Knoxville, TN, USA; ³⁷Department of Civil Engineering, Construction Management and Environmental Engineering, Northern Arizona University, Flagstaff, AZ, USA; ³⁸UCIBIO, REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal; ³⁹CSIRO Land and Water, Ecosciences Precinct, QLD, Australia; ⁴⁰Biodesign Swette Center for Environmental Biotechnology, Arizona State University, Tempe, AZ, USA; ⁴¹Departamento de Química, Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto – FFCLRP, Ribeirão Preto, SP, Brazil; ⁴²Department of Civil, Construction, and Environmental Engineering, North Carolina State University, NC, USA; ⁴³Grupo de Investigación en Procesos Anaerobios, Instituto de Ingeniería, Universidad Nacional Autónoma de México, México, DF, Mexico; ⁴⁴CNR-IRSA, National Research Council, Water Research Institute, Rome, Italy; ⁴⁵Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow, UK; ⁴⁶Tryon Creek & Columbia Blvd. Wastewater Treatment Plants, Bureau of Environmental Services, City of Portland, OR, USA; ⁴⁷Scion Research, Christchurch, New Zealand; ⁴⁸School of Engineering, University of Guelph, Guelph, Ontario, Canada; ⁴⁹Department of Environmental Engineering, National Cheng Kung University, Tainan City, Taiwan, ROC; ⁵⁰State Key Laboratory of Applied Microbiology Southern China, Guangdong Provincial Key Laboratory of Microbial Culture Collection and Application, Guangdong Institute of Microbiology, Guangzhou, Guangdong, China; ⁵¹Department of Civil and Environmental Engineering, University of Hawaii at Manoa, Honolulu, HI, USA; ⁵²State Key Laboratory of Environmental Aquatic Chemistry, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China; ⁵³State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China; ⁵⁴Department of Civil Engineering, University of Arkansas, Fayetteville, AR, USA; ⁵⁵Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; ⁵⁶School of Engineering, Newcastle University, Newcastle upon Tyne, UK; ⁵⁷Institute for a Secure and Sustainable Environment, The University of Tennessee, Knoxville, TN, USA; ⁵⁸Environmental Microbiomics Research Center, School of Environmental Science and Engineering, Sun Yat-Sen University, Guangzhou, China; ⁵⁹Guangdong Provincial Key Laboratory of Environmental Pollution Control and Remediation Technology, Guangzhou, China; ⁶⁰Consolidated Core Laboratory, University of Oklahoma, Norman, Oklahoma, USA.

Competing interests

The authors declare no competing financial interests.

Methods

Global sampling and meta-data collection

The Global Water Microbiome Consortium (GWMC) was initiated in May 2014 as a platform to facilitate international collaboration and communication on research and education for global water microbiome studies (<http://gwmc.ou.edu/>). The GWMC is a collaboration across more than 70 research groups from 23 countries. As the first initiative of GWMC, we launched this study with a global sampling campaign targeting municipal wastewater treatment plants (WWTPs) by focusing on the activated sludge process. Unlike the Earth Microbiome Project (EMP), which employed a bottom-up strategy to solicit microbial samples⁵, we used a top-down approach to select WWTPs for sampling by considering their latitudes, climate zones, spatial scales, activated sludge process type, and accessibility for sampling.

The main goal of this study was to provide system-level mechanistic understanding of global diversity and distribution of municipal WWTP microbiomes. WWTPs were selected based on the following criteria: **(i) Continental-level geographic locations.** Samples were obtained from all continents except for Antarctica, but with special focus on North America, Asia, and Europe (Fig. 1a). Because of the low accessibility, WWTPs in Africa and South America were under-represented. **(ii) Latitude.** To address questions related to latitudinal diversity gradient (LDG), WWTPs were intensively sampled in North America along the East and West Coasts, and Highway 35, as well as Highway 40 (from East to West) (Fig. 1a), in Asia, Europe, and Australia. The WWTPs sampled spanned latitudes from 43.6°S to 64.8°N. **(iii) Climate zones.** Since climate could have significant impacts on microbial communities, the samples covered 17

different climate types (Supplementary Fig. 6). To distinguish independent effects of continents versus climate zones, we increased sampling efforts for climate zones that were present in multiple continents, such as Humid Subtropical Climate. **(iv) Scales.** The samples were collected from very broad spatial scales: global (across 6 continents), regional (e.g., individual continents or climate zones), and local (e.g., individual cities). Within some cities, multiple WWTPs and multiple samples per WWTP were collected; **(v) Wastewater treatment process types.** To address the relationship of structure to function for activated sludge, we sampled the aerobic zone of conventional plug flow, oxidation ditch, sequential batch reactors, anaerobic/anoxic/oxic (A^2O), and other activated sludge process types.

A unified protocol was used for sampling, sample preservation, metadata collection, DNA extraction, sequencing, and sequence analysis, to minimize potential experimental variations^{4,51-53}. Detailed sampling and metadata collection methods and protocols are available at the GWMC web site (<http://gwmc.ou.edu/protocols/view/11>).

Sampling was carried out in June to November 2014 in the Northern Hemisphere and December 2014 to April 2015 in the Southern Hemisphere. The sampling time was generally between 10:00 am to 2:00 pm, when the WWTPs were relatively stable under normal conditions. Although we tried to collect the global samples in the same season, seasonal temporal turnover in activated sludge communities could have had some effect on the community variations we observed to some degree. Based on limited published work^{54,55}, such temporal variations should be much smaller than the spatial variations at the global/continental scales. For example, a previous study on 5-year temporal dynamics of activated sludge community showed no

significant seasonal succession⁵⁴. It's also revealed that the activated sludge communities were relatively stable across three months, with average Bray-Curtis distance 0.45 ± 0.10 (mean \pm SD) between samples⁵⁵; this variation was smaller than our observed mean variations even at local city level (0.54 ± 0.19) (Fig. 4a).

At local scale, we defined a city based on it having a large enough geographic scale, not on an administrative division (see Supplementary Table 1 for defined cities). For each city, we usually collected at least 12 samples, and had ≥ 12 samples/city in 77% cities, with < 3 samples/city in only 1% of cities. We also sampled at least 2 WWTPs in 72% of the cities. In each plant, we collected at least 3 mixed liquor samples, generally from 3 different positions (the front, middle, and end part) of the aerobic zone in each aeration tank. In a few cases (3.3% plants), where only one sampling position was applicable, 3 samples were taken in sequence with at least 30-min interval. Altogether, we collected 1,186 activated sludge samples from 269 WWTPs across 23 countries from global scale (e.g., across 6 continents), regional scale (e.g., individual continents), to local scale (e.g., geographic sites or individual cities) (Fig. 1a).

At each sampling position, approximately 1 liter mixed liquor was sampled and well mixed, and 40 mL was transferred into a sterile tube. The mixed liquor samples were kept on ice ($\leq 4^{\circ}\text{C}$), transported to laboratory within 24 hours, divided into aliquots, and then centrifuged at 4°C , 15,000 g for 10 min to collect pellets. Sludge pellets were transported (if necessary) with dry ice to the designated laboratories within 48 hours and preserved at -80°C before DNA extraction.

Along with the sludge samples, associated metadata, conforming to the Genomic Standards Consortium's MIxS and Environmental Ontology (ENVO) Standards^{56,57}, were provided by plant managers and/or investigators (Supplementary Table 1; Supplementary Fig. 7). We collected metadata (e.g., chemical properties, operation conditions, process type) from each plant using a standard sampling data sheet, which ensured that the data from all plants was in the same format. Raw metadata were processed as one metadata table (Supplementary Table 1) and classified into three categories: geological variables, plant operation and monitoring variables, and sample properties. The geological variables included latitude and longitude; ambient climate variables such as climate type, mean annual temperature (MAT), and precipitation; and population size and gross domestic product (GDP) for the city where the WWTP was located.

Climate type was determined by the Köppen-Geiger climate classification⁵⁸. GDP and population data were derived from the Brookings analysis of Global Metro Monitor⁵⁹. Variables related to plant design and operation include plant age, design capacity, actual flow rate, volume of aeration tanks, hydraulic retention time (HRT) and solids retention time (SRT). The activated sludge process type, aerator type, and coupling with N removal processes (nitrification and denitrification) in the WWTP were also provided by the plant managers as possible. Plant monitoring variables include influent and effluent biochemical oxygen demand (BOD) and chemical oxygen demand (COD) representing organic carbon (C) level, total nitrogen and total phosphorus representing nutrient level, ammonium N, as well as the food to microorganism (F/M) ratio, indicating the average organic C loading to microorganisms. For sample properties, most plant managers provided the yearly average value of mixed liquor suspended solids (MLSS),

indicating the concentration of biomass in the activated sludge, dissolved oxygen, pH, and mixed liquid temperature; some provided the measured values when sampling.

Activated sludge performance was calculated as the specific removal rates (g per g biomass per day) of organic C (BOD and COD), nutrients (total nitrogen and total phosphorus) and ammonium nitrogen (NH₄-N):

$$\text{removal rate} = \frac{(\text{Influent}(X) - \text{Effluent}(X)) \times \text{flow rate}}{\text{MLSS} \times \text{aerobic tank volume}}$$

The WWTPs represent diverse geographies and a large range of climatic conditions, operation parameters, and chemical conditions across and within continents (Supplementary Fig. 7). For instance, the average influent BOD ranged from 30 to 1,000 mg/L. Such a broad range of diverse parameters is critical to disentangling mechanisms of activated sludge microbial community assembly.

DNA Extraction

To minimize the variations associated with sample processing, identical protocols were used in DNA extraction and 16S rRNA gene sequencing. All samples from China and Japan were shipped to Dr. Xianghua Wen's Laboratory at Tsinghua University for DNA extraction. All other samples, including samples from Europe collected by Dr. Thomas Curtis at Newcastle University, were shipped to Dr. Jizhong Zhou's Laboratory at University of Oklahoma (OU) for

DNA extraction. Due to the tight restriction of sample shipment in South Africa, Mexico, Chile, Uruguay, and Brazil, the DNA was extracted by GWMC members in these countries. DNA was extracted from sludge samples using MoBio PowerSoil DNA isolation kit. For each sample, a pellet from 3 mL mixed liquor was used. In addition to the manufacture protocol, we always placed exactly 12 bead tubes on the vortex evenly and vortex at maximum speed for 10 min to minimize the lysis efficiency difference between samples. All DNA samples were processed at OU for sequencing.

DNA quality for all samples was evaluated with a NanoDrop spectrophotometer (NanoDrop Technologies Inc., Wilmington, DE, USA) at OU. Final DNA concentrations were quantified using PicoGreen with a FLUO star Optima instrument (BMG Labtech, Jena, Germany). Purified DNA was stored at -80 °C.

16S rRNA gene sequencing and sequence processing

The V4 region of the 16S rRNA gene was amplified and sequenced using standardized protocols with the phasing amplicon sequencing (PAS) approach as described previously⁶⁰ and the primers 515F (GTGCCAGCMGCCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) of the Earth Microbiome Project⁶¹. *In silico* primer coverage analysis using SILVA TestPrime 1.0⁶² and SILVA dataset r123 showed that these primers cover 86.8% and 52.9% of all bacterial and archaeal sequences with 0 mismatches, respectively.

To mitigate quantitative problems associated with amplicon sequencing⁵², the 16S rRNA gene fragments were amplified from community DNAs (10 ng) with two-step PCR using lower numbers of amplification cycles (10 and 20 cycles for the 1st and 2nd step, respectively). The two-step PAS approach offers several advantages: lower amplification biases, better sequence-read quality, higher effective sequence read numbers and length, and lower sequencing errors⁶⁰. All samples were sequenced using the same MiSeq instrument at the Institute for Environmental Genomics, OU. Generally, about 400 samples were combined for each round of MiSeq sequencing. Since the numbers of sequence reads varied substantially from sample to sample, most samples were sequenced more than once (e.g., 19% twice; 33%, three times; 43%, > 3 times) to meet the target number of about 30K sequencing reads per sample, as determined in our previous analysis⁶³.

The numbers of sequences (reads) per sample ranged from 25,631 to 351,844 (Supplementary Table 5), and a total of 96,148 OTUs were obtained. About 1.3% of these OTUs were from archaea, which accounted for 0.13% of the total abundance. The choice of the PCR primer pair 506F/806R (that was also used in the EMP project) is very likely to have strongly influenced this low archaeal abundance due to the much lower coverage of the primers of archaeal 16S rRNA genes compared to the bacterial counterparts. Because of the low archaeal abundance, the term “bacteria” is used for simplicity. Also, the terms microbiome and microbial (or bacterial) community are used interchangeably.

Raw sequence data were processed as previously described³⁵, except for OTU generation by UPARSE⁶⁴ at the 97% similarity threshold, resulting in 96,148 OTUs. We define operational

taxonomic units (OTUs) (based on 97% sequence similarity) for bacterial and archaeal
phylotypes. Although there is potential misconnection between OTUs and microbial species⁶⁵,
we use this popular definition for simplicity, and it also allows us to compare with previous
studies of other systems. The representative sequences were aligned using Clustal Omega
v1.2.2⁶⁶ for constructing the phylogenetic tree by FastTree2 v2.1.10⁶⁷. OTUs were
taxonomically annotated by RDP Classifier⁶⁸ with a confidence cutoff of 80%, using the MiDAS
database (version 2.1) which specifically provides a curated taxonomy for abundant and
functionally important microorganisms in activated sludge⁶⁹. After removal of the global
singletons⁶⁴, the sequence number in each sample was rarefied to the same depth (25,600
sequences per sample), resulting in 61,448 OTUs overall, which were used in subsequent
comparative analyses.

Although our sequencing depths were considerably higher than those in many similar studies⁷⁰,
rarefaction curves (Supplementary Fig. 2d, e) of activated sludge microbial communities
indicated that additional rare taxa were likely present in individual samples. Nevertheless,
pooling all sequences gave a sufficient number for estimating global- and continent-level
diversity of activated sludge microbial communities (Supplementary Fig. 2f, g). The global
OTU richness per sample was 2,309±559 (Supplementary Table 5). Besides richness, we also
calculated other alpha diversity indices on a global and regional scale (Supplementary Table 5).

The rRNA operon copy number for each OTU was estimated through the rrnDB database based
on its closest relatives with known rRNA operon copy number⁷¹. The abundance-weighted mean
rRNA operon copy number was then calculated for each sample as described previously⁴⁹.

Sequence comparison against reference databases

To compare the sequence diversity in this study to that in existing databases, the 96,148 representative sequences from the activated sludge samples were compared against the representative set (97% similarity level) of full-length sequences from Greengenes 13.8⁷² (released on August 2013) and the non-eukaryotic fraction of Silva 132 databases⁷³ (released on December 2017). We used the open-source sequence search tool USEARCH10⁷⁴ in global alignment search mode, and we required 97% similarity across the query sequence. Our activated sludge sequences match to 38.6% of Greengenes and 37.2% of SILVA 16S rRNA gene OTUs at 97% similarity. These matches accounted for 18.2% and 22.5% of the representative sequences in our datasets, respectively, indicating that the majority of activated sludge microbial species diversity is not yet captured in full-length sequence databases; this is similar to the observations in the EMP⁵.

Species abundance distribution (SAD) fitting

We compared the SAD of each sample, based on the rank-abundance distribution, with predictions from Poisson lognormal, log-series, Broken-stick, and Zipf models. Although numerous SAD models are available, lognormal and log-series have been the most successful in predicting SADs, and they are the standards for testing other models¹⁸. While the logseries model is well supported by macroecological studies, the Poisson lognormal model is more commonly observed with microorganisms¹⁸. By comparing (rank-for-rank) the observed and

predicted SADs using regression analysis, we could directly infer the percentages of variations in abundance among species explained by each model using the same code, developed by Shoemaker et al¹⁸.

Estimation of global bacterial diversity of WWTPs

We used the methods described in Curtis et al.²⁰ and Locey and Lennon¹⁹ to predict global bacterial richness (S_T) using the lognormal model. The lognormal prediction of S_T is based on the total abundance (N_T), the abundance of the most abundant species (N_{max}), and the assumption that the rarest species is a singleton, $N_{min} = 1$. In communities with N_T individuals, the richness can be estimated by:

$$S_T = \frac{\sqrt{\pi}}{a} \exp \left\{ \left(a \log_2 \left(\sqrt{\frac{N_{max}}{N_{min}}} \right) \right)^2 \right\} \quad (i)$$

where a is an inverse measure of the width of the distribution, which can be numerically solved from:

$$N_T = \frac{\sqrt{\pi N_{min} N_{max}}}{2a} \exp \left\{ \left(a \log_2 \left(\sqrt{\frac{N_{max}}{N_{min}}} \right) \right)^2 \right\} \exp \left\{ \left(\frac{\ln(2)}{2a} \right)^2 \right\} \left[\operatorname{erf} \left(a \log_2 \left(\sqrt{\frac{N_{max}}{N_{min}}} - \frac{\ln(2)}{2a} \right) \right) + \operatorname{erf} \left(a \log_2 \left(\sqrt{\frac{N_{max}}{N_{min}}} + \frac{\ln(2)}{2a} \right) \right) \right] \quad (ii)$$

We used published data to estimate the total microbial abundance in WWTPs as follows. Empirical records compiled from a variety of sources, for example, AQUASTAT⁷⁵ and Sato et al 2013⁷⁶, suggest that about 330 km³ year⁻¹ of municipal wastewater are produced globally, of

869 which 60% is treated⁹. Assuming that they are all treated in WWTPs, then about 0.54 km³
870 municipal wastewater are treated by WWTPs globally per day. The total effective volume of
871 aerobic tanks of WWTPs can be estimated by:

$$872 \quad V = Q \times HRT \quad (iii)$$

873 where Q is the influent flow rate (m³ day⁻¹) and HRT is the hydraulic retention time (day) of the
874 aerobic tank. Our dataset indicates that the average HRT of aerobic tanks is 9.8 (± 0.3 s.e.) hours.
875 Thus, the total effective volume is estimated as 0.22 (± 0.007) km³. The total cells in activated
876 sludge are about $2.3 (\pm 0.4) \times 10^9$ (ml⁻¹)⁷⁷; thus, N_T (global activated sludge bacterial abundance)
877 is about $4.0\text{--}6.1 \times 10^{23}$.

878

879 We then estimated N_{\max} based on the ratio of N_{\max} to N_T of our sequencing data, i.e., the relative
880 abundance of the most abundant OTU, or using scaling law¹⁹. The knowledge of N_T , N_{\max} , and
881 N_{\min} allows equation (ii) to be solved numerically for the parameter a and, subsequently, for S_T
882 using equation (i).

883

884 Using the same method, we estimated the total bacterial richness of individual WWTPs, along
885 with WWTPs in the United States and China. The volume of aerobic tanks of a WWTP in
886 Beijing, China is 10,000 m³, making the total cells about $2.3 (\pm 0.4) \times 10^{19}$. N_T of WWTPs in
887 US and China were estimated based on their published data of treating amount^{78,79}, with
888 activated sludge harboring similar numbers of species for the US (4.6×10^8 to 1.1×10^9) and
889 China (3.9×10^8 to 1.0×10^9). N_{\max} was further estimated based on our 16S rRNA gene
890 sequencing data or using a scaling law¹⁹. The total bacterial richness estimates of individual
891 human gut, individual cow rumen, global ocean and Earth were taken from Locey and Lennon¹⁹.

Core community determination

A global-scale core microbial community was determined based on multiple reported measures. First, “overall abundant OTUs” were filtered out according to mean relative abundance across all samples (MRA)⁸⁰. Previous studies used different criteria (e.g., MRA > 1%^{30,81} or 0.1%^{82,83}) without any objective or standard rule. Thus, we selected all top 0.1% OTUs (62) as overall abundant OTUs. Their MRA was higher than 0.2%, within the range of reported criteria. Second, “ubiquitous OTUs” were defined as OTUs with occurrence frequency in more than 80% of all samples⁸⁴. Finally, “frequently abundant OTUs” were selected based on their relative abundances with a sample. In each sample, the OTUs were defined as abundant when they had a higher relative abundance than other OTUs and made up the top 80% of the reads in the sample¹⁴. A frequently abundant OTU was defined as abundant in at least half samples, which is stricter than the reported criterion (10 in 26 samples¹⁴). Since the above three measures are complementary to one another when defining core community, only OTUs fulfilling all three criteria were defined as the global scale core bacterial community.

Following the same criteria as described above, the core community was identified for each continent. That is, a core OTU for a specific continent should be one that was from the top 0.1% OTUs of that continent; a core OTU also had to be detected in more than 80% of the samples and dominant for more 50% of the samples of that continent.

Comparison of bacterial community composition of WWTPs to natural habitats and source tracking

915
916 We downloaded the OTU table of 16S rRNA gene amplicon studies from the EMP
917 ([ftp://ftp.microbio.me/emp/release1/otu_tables/closed_ref_greenegenes/emp_cr_gg_13_8.subset_](ftp://ftp.microbio.me/emp/release1/otu_tables/closed_ref_greenegenes/emp_cr_gg_13_8.subset_5k.biom)
918 [5k.biom](ftp://ftp.microbio.me/emp/release1/otu_tables/closed_ref_greenegenes/emp_cr_gg_13_8.subset_5k.biom))⁵. This table was generated using closed reference against Greengenes 13.8 and
919 contained 5,000 global samples from multiple habitats. To compare community compositions at
920 the OTU level, our activated sludge OTUs were repicked using closed reference against
921 Greengenes 13.8, which picked 68.1% of the sequences. This OTU table was then merged with
922 the EMP OTU table. To give relatively equal representation of samples across environments, we
923 further collapsed our activated sludge samples at the plant level by summing the abundance of
924 each OTU across samples of the same plant, resulting in 269 activated sludge samples. Our
925 activated sludge samples and the EMP samples from freshwater (including that from freshwater
926 and freshwater biofilm), ocean (including that from sea water and biofilm), animal feces, human
927 feces, soil and air were selected from the merged OTU table. We then subsampled to 10,000
928 sequences per sample. To compare microbial community compositions across habitats, the
929 Nonmetric Multidimensional Scaling (NMS) analysis was performed using the Bray-Curtis
930 dissimilarity matrix.

931
932 The proportion of each activated sludge microbiota attributable to freshwater, soil, ocean, animal
933 and human feces, and air at the genus level were estimated using SourceTracker³⁹, which was
934 run through QIIME with default settings using activated sludge microbiota as the sink and those
935 in other habitats as sources. Genera detected in less than 1% of the samples were filtered out
936 before source-tracking modeling.

937

Diversity analyses: α - and β -diversity and correlation with environment

Richness and Faith's index were used to measure taxonomic and phylogenetic α -diversity, respectively, and they were computed using the *Picante* R package⁸⁵. Other taxonomic α -diversity indices, including Shannon index, Simpson index and Pielou's evenness, were calculated using the *vegan* R package⁸⁶.

Bray-Curtis (abundance-based) and Sorensen (incidence-based) distances were calculated to represent the taxonomic β -diversity using the *vegan* R package⁸⁶. Canberra's distance was also calculated to give more weight to rare taxa, using the *vegan* R package⁸⁶. The weighted (abundance-based) and unweighted UniFrac (incidence-based) distance⁸⁷ were calculated to represent the phylogenetic β -diversity using the *GUniFrac* R package⁸⁸. For each environmental variable, we performed a partial Mantel test to examine the correlation between environmental variable and microbial community composition independent of geographical location (999 permutations) using the *vegan* R package⁸⁶.

PERMANOVA was applied to assess the difference of community composition among continents, climate types, and activated sludge process types using the *vegan* R package⁸⁶. In PERMANOVA, climate types were defined at main climate group level, which includes 5 groups: A (tropical), B (arid), C (temperate), D (cold), and E (polar)⁵⁸. The activated sludge process types were classified into 9 general groups: complete mix, conventional plug flow, sequential batch reactors (SBR), anaerobic/anoxic/oxic (A²O), anoxic/oxic (AO), oxidation ditch, contact stabilization, pure oxygen and extended aeration.

Distance Decay relationships

The rate of the distance-decay relationship (DDR) was calculated as the slope of a linear least squares regression on the relationship between ln-transformed geographic distance versus ln-transformed bacterial community composition similarity. We used matrix permutation tests to examine the statistical significance of the distance-decay slope³⁷. The samples were permuted 999 times, and the observed slope was compared with the distribution of values in the permuted datasets. We also tested whether the slopes of the distance-decay curve at the three spatial scales (0 to 100 km; 100 to 5,000 km; and 5,000 to 25,000 km) were significantly different from the slope of the overall distance-decay curve, using matrix permutations to compare the observed difference between slopes within the three spatial scales with the overall distance-decay slope to that over 999 permutations.

Estimating stochasticity of community assembly

We assessed community-assembly stochasticity with a null-model-based index. The Stochasticity ratio was described previously^{41,89}. Since null-model algorithms usually require a high number of replicates, we selected 71 cities, each of which had more than 9 samples; we randomly drew 9 samples from each city to make sampling even. We calculated stochasticity ratio using taxonomic and phylogenetic metrics. Whether using the Bray-Curtis (abundance-weighted) or Sorensen (unweighted) model, the stochasticity ratio was calculated based on typical null-model algorithms for taxonomic metrics^{90,91}. When using weighted and unweighted

Unifrac, the stochasticity ratio was calculated based on typical null-model algorithms for phylogenetic metrics^{91,92}. Samples within each city were considered sharing the same regional species pool in null model algorithms.

Partitioning the environment and distance effect

To give a quantification of relative contribution of the environment effect versus the distance effect on β -diversity, we performed a variation partition analysis (VPA) based on multiple regression on matrices (MRM). We used a modified MRM approach as described previously³⁷. Briefly, we first selected a non-redundant environmental variable set. The final set included temperature, precipitation, design capacity, SRT, dissolved oxygen, pH, and influent BOD. The highest correlation was between design capacity and SRT (Pearson' $r = -0.25$), and it indicated a low level of collinearity among these variables. MRM was performed in different spatial scales. Geographic distance and microbial community distance were ln-transformed. A Euclidean distance matrix was calculated for each environmental variable. To reduce the effect of spurious relationships between variables, we first ran the MRM test with all the variables in the non-redundant environmental variable set, removed the non-significant variables from this initial MRM test, and then reran the test³⁷. The significance of the partial regression was tested by matrix permutation for 999 times⁹³. In VPA, the R^2 of the selected environmental variables as independent matrices (R^2_E), geographical distance as independent matrix (R^2_G), and all matrices (R^2_T) were used to compute the four components of variations as described elsewhere⁹⁴: (i) pure environmental variation = $R^2_T - R^2_G$; (ii) pure geographical distance = $R^2_T - R^2_E$; (iii)

1006 spatially structured environmental variation = $R^2_G + R^2_E - R^2_T$; and (iv) unexplained variation
1007 = $1 - R^2_T$.

1008

1009 **Structural equation model (SEM)**

1010

1011 SEM was used to explore the direct and indirect relationships among environmental variables,
1012 bacterial communities, and activated sludge function. The community composition was
1013 represented by the first principal component (PC1) of Principal coordinate analysis (PCoA)
1014 based on Bray-Curtis distance. We first considered a full model that included all reasonable
1015 pathways, and then we sequentially eliminated non-significant pathways until we attained the
1016 final model whose pathways all were significant. To capture the quadratic correlation of SRT to
1017 diversity and BOD removal, we constructed a composite variable⁹⁴ of ‘SRT effect’ as a linear
1018 combination of SRT and the square of SRT (SRT.SQ). We used a χ^2 test and the root mean
1019 square error of approximation to evaluate the fit of model. The SEM-related analysis was
1020 performed using the *lavaan* R package⁹⁵.

1021

1022 **Random Forest models**

1023

1024 We applied a machine-learning model, random forest, to examine the strengths of the
1025 associations between environmental variable and compositional data, using the *randomForest* R
1026 package⁹⁶. We used OTUs as predictors and environmental variable as response data. To
1027 correct the potential spatial autocorrelation, we used OTU data at the plant level, by averaging
1028 the relative abundance of each OTU across samples of the same plant. OTUs which were

detected in at least 20% of all the plants and in all continents were used for modelling. We allowed a baseline model to learn using the full data-set for training, and subsequently, we trained new random forests for each plant with customized training sets that excluded plants within a defined radius of the target plant. The size of this radius ranged from 0 to 5000 km. To delineate the model prediction strength, the cross-validated R^2 was calculated as $1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y}_i)^2}$, where y_i is the value of the parameter for sample i , \hat{y}_i is the prediction for that same sample (obtained by held-out cross-validation), and \bar{y}_i is the overall mean (the summation runs over all the samples).

Data availability

The sample metadata are available in Supplementary Table 1. Sequences are available from the NCBI Sequence Read Archive with accession number PRJNA509305. OTU tables and representative sequences of the OTUs are available on the GWMC web site (<http://gwmc.ou.edu/data-disclose.html>).

Code availability

R codes on the statistical analyses are available at <https://github.com/Linwei-Wu/Global-bacterial-diversity-in-WWTps>.

References of Methods

- 51 Zhou, J. *et al.* Random Sampling Process Leads to Overestimation of β -Diversity of Microbial Communities. *mBio* **4** (2013).
- 52 Zhou, J. *et al.* Reproducibility and quantitation of amplicon sequencing-based detection. *ISME J* **5**, 1303-1313 (2011).

1056 53 Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by
1057 the Microbiome Quality Control (MBQC) project consortium. *Nat Biotechnol* **35**, 1077
1058 (2017).

1059 54 Ju, F. & Zhang, T. Bacterial assembly and temporal dynamics in activated sludge of a
1060 full-scale municipal wastewater treatment plant. *ISME J* **9**, 683-695 (2015).

1061 55 Xia, Yu. *Diversity and temporal assembly patterns of microbial communities in*
1062 *municipal wastewater treatment systems*. PhD thesis, Univ. Tsinghua, Beijing, China
1063 (2016).

1064 56 Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J. & Lewis, S. E. The environment
1065 ontology: contextualising biological and biomedical entities. *J Biomed Semantics* **4**, 43
1066 (2013).

1067 57 Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and
1068 minimum information about any (x) sequence (MIXS) specifications. *Nat Biotechnol* **29**,
1069 415-420 (2011).

1070 58 Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen-
1071 Geiger climate classification. *Hydrol Earth Syst Sci Discuss* **4**, 439-473 (2007).

1072 59 Alan Berube, J. L. T., Tao Ran, Joseph Parilla. *Global Metro Monitor*,
1073 <<https://www.brookings.edu/research/global-metro-monitor/>> (2015).

1074 60 Wu, L. *et al.* Phasing amplicon sequencing on Illumina Miseq for robust environmental
1075 microbial community analysis. *BMC Microbiol* **15**, 125 (2015).

1076 61 Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of
1077 sequences per sample. *Proc Natl Acad Sci USA* **108**, 4516-4522 (2011).

1078 62 Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for
1079 classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**,
1080 e1-e1 (2013).

1081 63 Wen, C. *et al.* Evaluation of the reproducibility of amplicon sequencing with Illumina
1082 MiSeq platform. *PLoS One* **12**, e0176716 (2017).

1083 64 Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads.
1084 *Nat Methods* **10**, 996-998 (2013).

1085 65 McLaren, M. R. & Callahan, B. J. In Nature, There Is Only Diversity. *mBio* **9**, e02149-
1086 02117 (2018).

1087 66 Sievers, F. *et al.* Fast, scalable generation of high-quality protein multiple sequence
1088 alignments using Clustal Omega. *Mol Syst Biol* **7**, doi:10.1038/msb.2011.75 (2011).

1089 67 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately maximum-
1090 likelihood trees for large alignments. *PLoS One* **5**, e9490,
1091 doi:10.1371/journal.pone.0009490 (2010).

1092 68 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid
1093 assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol*
1094 **73**, 5261-5267 (2007).

1095 69 McIlroy, S. J. *et al.* MiDAS 2.0: an ecosystem-specific taxonomy and online database for
1096 the organisms of wastewater treatment systems expanded for anaerobic digester groups.
1097 *Database* **2017** (2017).

1098 70 Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil.
1099 *Science* **359**, 320-325 (2018).

1100 71 Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. & Schmidt, T. M. rrnDB: improved
1101 tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation
1102 for future development. *Nucleic Acids Res*, gku1201 (2014).

1103 72 McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for
1104 ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**, 610 (2012).

1105 73 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data
1106 processing and web-based tools. *Nucleic Acids Res* **41**, D590-D596 (2012).

1107 74 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST.
1108 *Bioinformatics* **26**, 2460-2461 (2010).

1109 75 AQUASTAT. *FAO global information system on water and agriculture. Wastewater*
1110 *section.*, <<http://www.fao.org/nr/water/aquastat/wastewater/index.stm>> (2014).

1111 76 Sato, T., Qadir, M., Yamamoto, S., Endo, T. & Zahoor, A. Global, regional, and country
1112 level need for data on wastewater generation, treatment, and use. *Agri Water Manag* **130**,
1113 1-13 (2013).

1114 77 Foladori, P., Bruni, L., Tamburini, S. & Ziglio, G. Direct quantification of bacterial
1115 biomass in influent, effluent and activated sludge of wastewater treatment plants by using
1116 flow cytometry. *Water Res* **44**, 3807-3818 (2010).

1117 78 Agency, U. S. E. P. *The Sources and Solutions: Wastewater*,
1118 <<https://www.epa.gov/nutrientpollution/sources-and-solutions-wastewater>> (2018).

1119 79 Chan, W. *Wastewater: Good To The Last Drop*,
1120 <<http://chinawaterrisk.org/resources/analysis-reviews/wastewater-good-to-the-last-drop/>>
1121 (2017).

1122 80 Hanski, I. Dynamics of regional distribution: the core and satellite species hypothesis.
1123 *Oikos*, 210-221 (1982).

1124 81 Galand, P. E., Casamayor, E. O., Kirchman, D. L. & Lovejoy, C. Ecology of the rare
1125 microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci USA* **106**, 22427-22432,
1126 doi:10.1073/pnas.0908284106 (2009).

1127 82 Székely, A. J. & Langenheder, S. The importance of species sorting differs between
1128 habitat generalists and specialists in bacterial communities. *FEMS Microbiol Ecol* **87**,
1129 102-112 (2014).

1130 83 Cheng, J. *et al.* Discordant temporal development of bacterial phyla and the emergence of
1131 core in the fecal microbiota of young children. *ISME J* **10**, 1002 (2016).

1132 84 Ju, F. & Zhang, T. Bacterial assembly and temporal dynamics in activated sludge of a
1133 full-scale municipal wastewater treatment plant. *ISME J*, doi:10.1038/ismej.2014.162
1134 (2014).

1135 85 Kembel, S. W. *et al.* Picante: R tools for integrating phylogenies and ecology.
1136 *Bioinformatics* **26**, 1463-1464 (2010).

1137 86 Oksanen, J. *et al.* Package ‘vegan’. *Community ecology package, version 2* (2013).

1138 87 Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial
1139 communities. *Appl Environ Microbiol* **71**, 8228-8235 (2005).

1140 88 Chen, J. GUniFrac: generalized UniFrac distances. *R package version 1*, 2012 (2012).

1141 89 Guo, X. *et al.* Climate warming leads to divergent succession of grassland microbial
1142 communities. *Nat Clim Change* **8**, 813 (2018).

1143 90 Chase, J. M., Kraft, N. J., Smith, K. G., Vellend, M. & Inouye, B. D. Using null models
1144 to disentangle variation in community dissimilarity from variation in α -diversity.
1145 *Ecosphere* **2**, art24 (2011).

1146 91 Stegen, J. C. *et al.* Quantifying community assembly processes and identifying features
1147 that impose them. *ISME J* **7**, 2069-2079 (2013).
1148 92 Kembel, S. W. Disentangling niche and neutral influences on community assembly:
1149 assessing the performance of community phylogenetic structure tests. *Ecol Lett* **12**, 949-
1150 960 (2009).
1151 93 Legendre, P., Lapointe, F. J. & Casgrain, P. Modeling brain evolution from behavior: a
1152 permutational regression approach. *Evolution* **48**, 1487-1499 (1994).
1153 94 Grace, J. B. & Bollen, K. A. Representing general theoretical concepts in structural
1154 equation models: the role of composite variables. *Environ Ecol Stat* **15**, 191-213 (2008).
1155 95 Rosseel, Y. Llavaan: An R package for structural equation modeling and more. Version
1156 0.5-12 (BETA). *Journal of statistical software* **48**, 1-36 (2012).
1157 96 Liaw, A. & Wiener, M. Classification and regression by randomForest. *R news* **2**, 18-22
1158 (2002).
1159 97 Faith, D. P. Conservation evaluation and phylogenetic diversity. *Biol Cons* **61**, 1-10
1160 (1992).
1161

Figures legends

Fig. 1. The Global Water Microbiome Consortium captures microbial diversity of globally distributed wastewater treatment plants (WWTPs). (a) Geographical distribution of 269 WWTPs where activated sludge samples and environmental data were collected. (b) Predicting species abundance distribution (SAD) of activated sludge bacterial communities. The grey line represents a SAD that was randomly chosen from our data. Each model was fit to the observed SAD (see Methods). Supplementary Fig.1a shows the variations of the SADs explained by each model across all 1186 activated sludge communities, indicating the best performance of the Poisson lognormal model. (c) Estimation of activated sludge microbial richness of WWTPs. Microbial species are defined as OTUs at 97% sequence similarity threshold. The microbial richness (S)-abundance (N) scaling relationship (dashed grey line with pink hull as 95% prediction interval), and the grey dots representing richness estimates from other systems were derived from Locey and Lennon¹⁹. Richness was predicted from the lognormal model using N_T estimated from published data, and N_{\max} inferred from our sequencing data (filled circle) or N_{\max} predicted from the dominance-scaling law¹⁹ (hollow circles). ‘WWTP’ indicates one WWTP, as do ‘Human gut’ and ‘Cow rumen’. (d) Latitudinal distribution of activated sludge bacterial diversity, plotting OTU richness against the absolute latitude of sampling locations shows the peak of richness at intermediate latitude ($n = 1,186$ biologically independent samples). The line shows the second order polynomial fit based on ordinary least squares regression. $P < 2 \times 10^{-16}$ (two-sided) for both regression coefficients. The color gradient denotes the annual mean air temperature. Shapes of symbols denotes whether a sample originated from Northern (circle) or Southern Hemisphere (square).

Fig. 2. Abundance, composition and functional importance of the global core operational taxonomic units (OTUs) in activated sludge. (a) Percentage and relative abundance of the global core OTUs versus the remaining microbial OTUs. In total, 0.05% (28 out of 61,448 OTUs) were identified as abundant and ubiquitous across wastewater treatment plants at global scale, which accounted for 12.4% of the 16S rRNA gene sequences in an activated sludge sample on average. (b) The taxonomic composition of the global core OTUs on phylum and class level. (c) Activated sludge functions were calculated as the removal rate of organic carbon (biochemical oxygen demand (BOD) removal, chemical oxygen demand (COD) removal), nutrients (total nitrogen (TN) and total phosphorus (TP) removal) and ammonia nitrogen ($\text{NH}_4\text{-N}$ removal) (g chemical per g MLSS per day, where MLSS is mixed liquor suspended solids relating to microbial biomass). The color gradient on the right indicates Spearman’s rank correlation coefficients, with more positive values (dark blue) indicating stronger positive correlations, and more negative values (dark red) indicating stronger negative correlations. The asterisks denote the significance levels (two-sided) of the Spearman’s rank correlation coefficients ($n = 1,186$ biologically independent samples): *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. In the correlation analysis, all OTUs detected in at least 20% of samples were included, and P values were adjusted for multiple testing using the Benjamini and Hochberg false

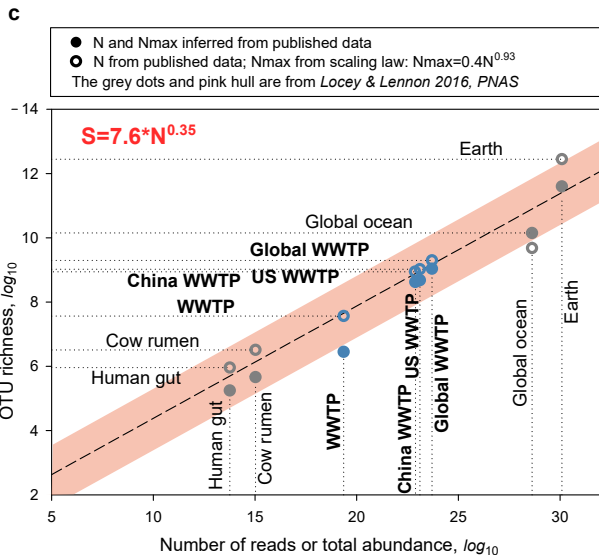
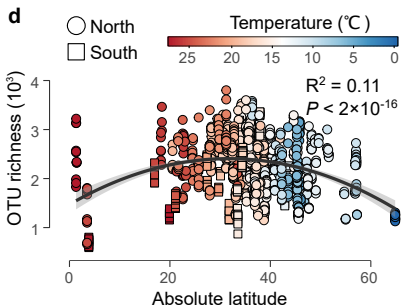
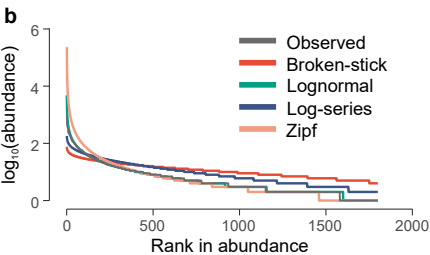
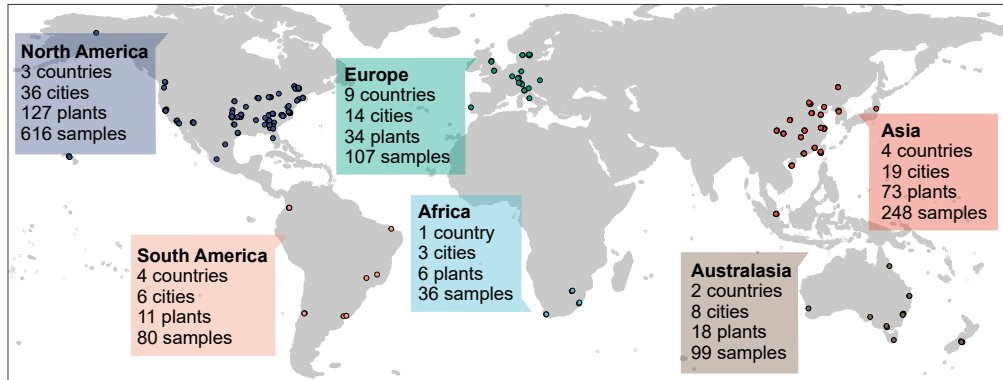
discovery rate (FDR) controlling procedure ($n = 14,235$ pairwise cases). Only global core OTUs were shown, with their mean relative abundance indicated on the left of the heatmap.

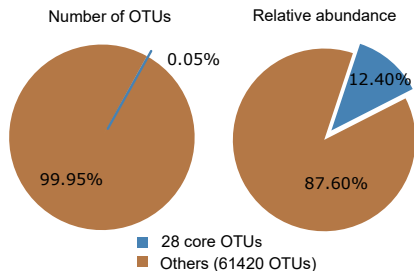
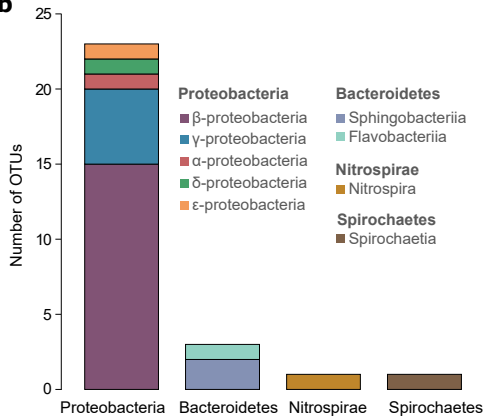
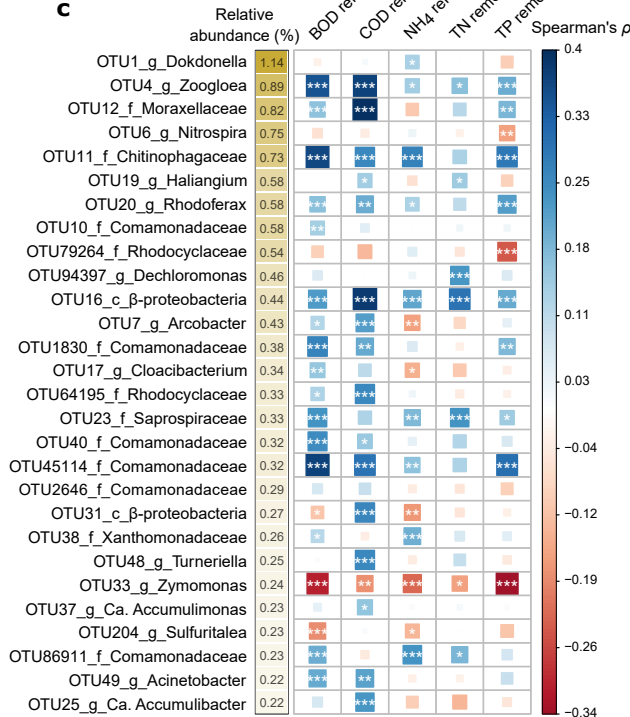
Fig. 3. Comparing bacterial community compositions across continents and with other habitats. (a) Nonmetric Multidimensional Scaling analysis (NMDS) showing that activated sludge of WWTPs harbored a unique microbiome as compared with other habitats. For comparison, we merged our OTU table ($n = 269$ WWTPs) with that released by EMP⁵, which contained thousands of bacterial communities from various habitats such as soil ($n = 338$ samples), ocean ($n = 969$ samples), freshwater ($n = 447$ samples), air ($n = 81$ samples), human feces ($n = 99$ samples) and animal feces ($n = 622$ samples), but not activated sludge from WWTPs (see Methods for details). Bray-Curtis distance was calculated to represent the dissimilarity in bacterial community compositions. (b) Percentage of activated sludge bacterial genera attributable to air, animal and human feces, freshwater, ocean and soil, as determined by SourceTracker. In the boxplots, hinges show the 25, 50 and 75 percentiles. The upper whisker extends to the largest value no further than $1.5 * \text{IQR}$ from the upper hinge, where IQR is the inter-quartile range between the 25% and 75% quartiles; The lower whisker extends to the smallest value at most $1.5 * \text{IQR}$ from the lower hinge. Sample size: $n = 6, 73, 18, 34, 127$ and 11 WWTPs for Africa, Asia, Australasia, Europe, North America and South America, respectively.

Fig. 4. Spatial turnover of the activated sludge bacterial communities. (a) Distance-decay relationships (DDR) based on Bray-Curtis similarity. Black line denotes the least-squares linear regression across all spatial scales ($n = 702,705$ pairwise distances). Colored lines denote separate regressions: within cities ($n = 9,753$ pairwise distances), within continents ($n = 220,136$ pairwise distances), and intercontinental ($n = 472,816$ pairwise distances). P values (one-sided) for regression slopes were determined by matrix permutation tests. (b) Ecological stochasticity in bacterial community assembly estimated by stochasticity ratio, which is calculated for each pair of samples ($n = 71$ cities) based on taxonomic diversity (Taxo., Bray-Curtis/Sorensen) and phylogenetic diversity (Phyl., Unifrac) weighted with abundance (Wt) or not (Uw). Boxes and whiskers indicate quartiles and triangles indicate mean values. (c) Variance partition analysis showing relative contributions of geographic distance (Geo) and environmental variables (ENV) to the community variations based on Bray-Curtis distance.

Fig. 5. Environmental drivers of the activated sludge community composition. (a) A structural equation model (SEM) shows relationships among environmental variables, community composition, and WWTP functioning. The composite variable of ‘SRT effect’ was constructed as a linear combination of solids retention time (SRT) and the square of SRT (SRT.SQ). F/M is the food to microorganisms ratio. The community composition is represented by the first principal component score (PC1) from the Bray-Curtis distance-based principal coordinate analysis. Blue and red arrows represent significant ($P < 0.05$) positive and negative pathways, respectively. Numbers near the pathway arrow indicate the standard path coefficients

(β). Arrow width is proportional to the strength of the relationship. R^2 represents the proportion of variance explained for every dependent variable. Model $\chi^2 = 13.92$, $df = 12$, $P = 0.31$, $n = 1,186$ biologically independent samples; root mean square error of approximation (RMSEA) = 0.012 with probability of a close fit = 1.00. **(b)** The average rRNA gene copy number of the community increased with the influent biochemical oxygen demand (BOD)/(1+recycle ratio) which approximates the influent BOD level of aerobic tank ($n = 641$ biologically independent samples). The P value (two-sided) denotes the significance of the slope of ordinary least squares regression. **(c)** The strength of association between taxonomic composition and temperature was tested by random forest ($n = 269$ WWTPs). The red diagonal shows the theoretical curve for perfect predictions. The inset shows a model trained on data from North and South America samples to predict the temperature in Asian samples ($n = 73$ WWTPs).



a**b****c**

a

● WWTP ● Air ● Animal feces ● Soil
● Freshwater ● Human feces ● Ocean

NMDS2

Stress: 0.19

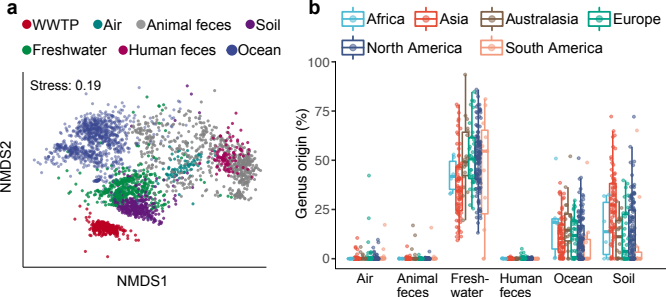
NMDS1

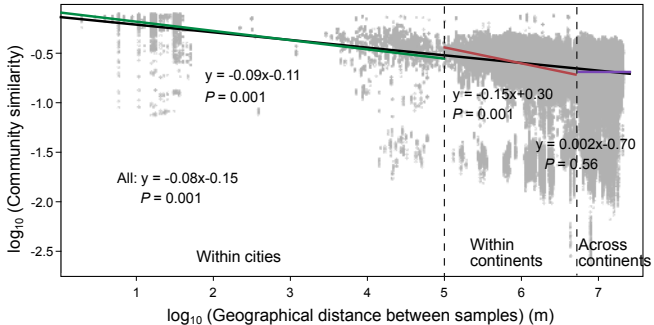
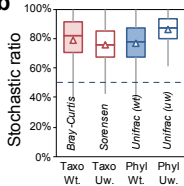
b

☐ Africa ☐ Asia ☐ Australasia ☐ Europe
☐ North America ☐ South America

Genus origin (%)

Air Animal feces Freshwater Human feces Ocean Soil



a**b****c**