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 Li[4], Ping Zhang[2,5], Xiaoyu Shan[1], Qiuting Zhang[1], Mathew Brown[6], Zhenxin Li[7], Joy D. Van Nostrand[2], Fangqiong Ling[8], Naijia Xiao[2,3], Ya Zhang[2], Julia Vierheilig[9,10], George F. Wells[11], Yunfeng Yang[1], Ye Deng[12,13], Qichao Tu[12], Aijie Wang[13], Global Water Microbiome Consortium‡, Tong Zhang[14], Zhili He[15,16], Jurg Keller[17], Per H. Nielsen[18], Pedro J. J. Alvarez[19], Craig S. Criddle[20], Michael Wagner[9], James M. Tiedje[21], Qiang He[22,23]*, Thomas P. Curtis[6]*, David A. Stahl[24], Lisa Alvarez-Cohen[25,26], Bruce E. Rittmann[27], Xianghua Wen[1]*, and Jizhong Zhou[2,1,26]*

[1]State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China; [2]Institute for Environmental Genomics, Department of Microbiology and Plant Biology, and School of Civil Engineering and Environmental Sciences, University of Oklahoma, Norman, OK, USA; [3]Consolidated Core Laboratory, University of Oklahoma, Norman, Oklahoma, USA; [4]College of Resource & Environment Southwest University, Chongqing, China; [5]Alkek Center for Metagenomics and Microbiome Research, Department of Molecular Virology and Microbiology, Baylor College of Medicine, Houston, TX, USA; [6]School of Engineering, Newcastle University, Newcastle upon Tyne, UK; [7]School of Environment, Northeastern Normal University, Changchun, China; [8]Department of Energy, Environmental and Chemical Engineering, Washington University in St. Louis, MO, USA; [9]Department of Microbiology and Ecosystem Science, Division of Microbial Ecology, Research Network 'Chemistry meets Microbiology', University of Vienna, Vienna, Austria; [10]Karl Landsteiner University of Health Sciences, Division of Water Quality and Health, Krems, Austria & Interuniversity Cooperation Centre for Water and Health; [11]Department of Civil and Environmental Engineering, Northwestern University, Evanston, IL, USA; [12]Institute for Marine Science and Technology, Shandong University, Qingdao, China; [13]Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China; [14]Environmental Biotechnology Laboratory, The University of Hong Kong, Hong Kong, China; [15]Environmental Microbiomics Research Center, School of Environmental Science and Engineering, Sun Yat-Sen University, Guangzhou, China; [16]Guangdong Provincial Key Laboratory of Environmental Pollution Control and Remediation Technology, Guangzhou, China; [17]Advanced Water Management Centre, The University of Queensland, Brisbane, QLD, Australia; [18]Department of Chemistry and Bioscience, Center for Microbial Communities, Aalborg University, Aalborg, Denmark; [19]Department of Civil and Environmental Engineering, Rice University, Houston, TX, USA; [20]Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA; [21]Center for Microbial Ecology, Michigan State University, East Lansing, MI, USA; [22]Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN, USA; [23]Institute for a Secure and Sustainable Environment, The University of Tennessee, Knoxville, TN, USA; [24]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA; [25]Department of Civil and Environmental Engineering, College of Engineering, University of California, Berkeley, CA, USA; [26]Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [27]Biodesign 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 Earth[1], 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 changes[2] and improving bioprocesses, such as wastewater treatment and soil remediation[3]. With recent advances in metagenomic technologies[4], microbial biodiversity and distribution are being intensively studied in a wide variety of environments[5-7], 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 km^3 of wastewater is produced globally each year[9]. This volume equals one seventh of the global river volume[10]. About 60% of this wastewater is treated prior to release, and biological processes such as activated sludge are widely used in WWTPs[9]. Activated sludge employs microbial flocs or granules to remove C, N, P, micropollutants (e.g., toxins, pesticides, hormones, pharmaceuticals), and pathogens[11]. Activated sludge relies on complex and incompletely defined microbial communities. As the largest application of biotechnology in the world[12], activated sludge is a vital infrastructure of modern urban societies[13]. Despite recent advances in understanding the microbial ecology of activated sludge[14-16], 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 sludge[3]. 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 ecology[17] and central to biodiversity theory, has not been rigorously tested in microbial ecology until recently[18]. 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 studies[18], 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 system[19]. 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 × 10^{23} bacteria in the global activated sludge community, based on published data[9]. 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 law[19]. The lognormal model predicts 1.1 (± 0.07) × 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 (± 0.2) × 10^9 species, using $N_{max} = 0.4 \times N_T^{0.93}$ from the dominance-scaling law[19] (Fig. 1c). The estimates of global activated sludge bacterial richness are only about one order of magnitude lower than that of the global ocean microbiome[19] (~ 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 Danish[14] and Asian[15] WWTPs, but less than 10% of the genera overlapped. Thus, a global core cannot be established from those regional studies.

142 At the global scale, occupancy-frequency and occupancy-abundance analyses revealed a hyper-
143 dominant pattern (Supplementary Fig. 1a) in which the 866 most abundant OTUs (1.39% of the
144 total OTU number) accounted for 50.06% of the total abundance. Similar hyper-dominance
145 patterns were observed in other macro-[21] and microbial communities[22].
146
147 A core bacterial community was determined based on abundance and occurrence frequency of
148 OTUs (see Methods for details). About 0.05% (28 OTUs) constituted a global core that
149 accounted for 12.4% ± 0.2% (mean ± SE) of the sequences in activated sludge samples (Fig. 2a;
150 Supplementary Table 3). Most (82%) of the core community members belonged to
151 *Proteobacteria*, with 15 OTUs classified as *β-Proteobacteria* (Fig. 2b). The most abundant OTU,
152 accounting for 1.14% ± 0.05% of the sequence abundance in activated sludge samples and
153 occurring in 85% of all samples, was 99% similar to the γ-proteobacterium *Dokdonella*
154 *kunshanensis DC-3* [23]. The second most abundant OTU (0.89% ± 0.06% in relative abundance
155 and occurring in 96% of all samples) belonged to *Zoogloea*, a dominant genus in activated
156 sludge communities[15], with *Z. ramigera* known to enhance the flocculation of activated sludge[24].
157 A *Nitrospira* OTU (OTU_6) was also identified as a core taxon, reflecting its importance for
158 nitrite oxidation or complete ammonia oxidation in activated sludge[25,26]. OTU_7 is closely
159 related to *Arcobacter* species, which are highly abundant in raw sewage[27] and include potential
160 pathogens, such as *A. cryaerophilus*, *A. butzleri*, and *A. skirrowii*[28]. Furthermore, two putative
161 polyphosphate- accumulating organisms (PAOs), a "*Candidatus* Accumulimonas" OTU
162 (OTU_37) and a "*Candidatus* Accumulibacter" OTU (OTU_25), were identified as core taxa,
163 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 systems[29]. However, some discrepancies also occurred. Saunders et al. showed *Nitrotoga* rather than *Nitrospira* as primary nitrite-oxidizers in Danish WWTPs[14]. 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 ratios[30]. 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) datasets[5]. 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 increases[32], are well documented in plant and animal ecology[33]. 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 ~15^{o}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 microbiome[7], but contrasts with some ocean[34] and soil communities[35]. In addition, the relationship between bacterial richness and temperature (Supplementary Fig. 2b, c) did not fit predictions from the metabolic theory of ecology[36]. 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 functioning[38]. 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 approach[39] 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 domains[37], 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 soils[41].

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 soils[6] and sediments[37].

Relationships between the community structure and activated sluge functions

Understanding the relationships between biodiversity and ecosystem function is a critical topic in ecology[42]. Despite decades of intensive studies, the biodiversity-function relationship is still hotly debated, particularly in microbial ecology[43]. 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 tests[44]. 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

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

287

288 Consistent with our expectation, the activated sludge community composition was significantly
289 correlated with the TP removal rate for the samples from EBPR plants, but not for non-EBPR
290 plants (Supplementary Table 10), as P removal processes in non-EBPR plants are predominantly
291 chemical. The diversity of the three potential PAOs[31] were significantly different ($P < 0.0001$,
292 two tailed paired-t test between any two organisms): 8.2 ± 0.2 "*Candidatus* Accumulimonas"
293 OTUs, 6.6 ± 0.2 "*Candidatus* Accumulibacter" OTUs, and 3.2 ± 0.1 *Tetrasphaera* OTUs within
294 a typical activated sludge bioreactor. While the relative abundance of "*Candidatus*
295 Accumulimonas" (0.42% ± 0.06%) was not different from that of "*Candidatus* Accumulibacter"
296 (0.42% ± 0.04%) (two tailed paired-t test, $P = 0.92$), both were more abundant than *Tetrasphaera*
297 (mean relative abundance 0.17% ± 0.02%) (two tailed paired-t test, $P < 0.0001$) (Supplementary
298 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 ratios[41] 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, β = 0.50, P < 0.001). It also had weak negative impacts on species richness (β = -0.14, P < 0.001). This is consistent with previous observations at local[45,46] and regional[47] 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

344 effect on the F/M ratio ($\beta = 0.31$, $P < 0.001$), which is inversely related to the SRT. Influent
345 BOD is the most influential environmental variable directly related to bacterial richness ($\beta = -$
346 0.28, $P < 0.001$), and the abundance-weighted mean rRNA gene copy number significantly
347 increased with the influent BOD ($R^2 = 0.19$, $P < 0.0001$; Fig. 5b). All of these results are
348 consistent with the resource-competition theory[48], which predicts that high species diversity
349 occurs with low to intermediate supply of resources, but fast-growing r-strategists outcompete
350 efficient-scavenging K-strategists at high resource levels[49].
351
352 To independently test the strength of correlation for each of the three strongest parameters
353 (temperature, SRT, and influent BOD) with bacterial community structure, we performed
354 random-forest analysis, a machine learning-based method. Using species abundance as the input
355 data, the model predicted temperature, SRT, and influent BOD with an explained variance of
356 69%, 25%, and 18%, respectively (Fig. 5c, Supplementary Fig. 5b). When controlling for spatial
357 auto-correlation, models of temperature continued to have higher accuracy (Supplementary Fig.
358 5b). For example, the America-fitted model of temperature, i.e., a model trained solely by North
359 and South America samples, was able to capture variations in the temperatures of Asia samples
360 (cross-validated $R^2 = 0.47$) (Fig. 5c). The random-forest model also revealed the most important
361 OTUs for predicting temperature (Supplementary Fig. 5c). These results corroborate that
362 temperature is the major environmental variable shaping the activated sludge bacterial
363 compositions at the global scale, although it only has a weak effect on species richness (Fig. 5a).
364

365 **Conclusions and future perspectives**

366

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 ~ 10^9 different bacterial species. In contrast, only about 10^4 species have been cultivated and studied in detail[19]. 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 engineers[50]. 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 Ofiţeru, 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).

23 Li, Y. *et al.* Dokdonella kunshanensis sp. nov., isolated from activated sludge, and emended description of the genus Dokdonella. *Int J Syst Evol Microbiol* **63**, 1519-1523 (2013).
24 Rosselló-Mora, R. A., Wagner, M., Amann, R. & Schleifer, K.-H. The abundance of Zoogloea ramigera in sewage treatment plants. *Appl Environ Microbiol* **61**, 702-707 (1995).
25 Daims, H. *et al.* Complete nitrification by Nitrospira bacteria. *Nature* **528**, 504 (2015).
26 Daims, H., Nielsen, J. L., Nielsen, P. H., Schleifer, K.-H. & Wagner, M. In Situ Characterization ofNitrospira-Like Nitrite-Oxidizing Bacteria Active in Wastewater Treatment Plants. *Appl Environ Microbiol* **67**, 5273-5284 (2001).
27 Fisher, J. C., Levican, A., Figueras, M. J. & McLellan, S. L. Population dynamics and ecology of Arcobacter in sewage. *Front Microbiol* **5** (2014).
28 Collado, L. & Figueras, M. J. Taxonomy, epidemiology, and clinical relevance of the genus Arcobacter. *Clin Microbiol Rev* **24**, 174-192 (2011).
29 Nielsen, P. H., Saunders, A. M., Hansen, A. A., Larsen, P. & Nielsen, J. L. Microbial communities involved in enhanced biological phosphorus removal from wastewater—a model system in environmental biotechnology. *Curr Opin Biotechnol* **23**, 452-459 (2012).
30 Lawson, C. E. *et al.* Rare taxa have potential to make metabolic contributions in enhanced biological phosphorus removal ecosystems. *Environ Microbiol* **17**, 4979-4993 (2015).
31 Stokholm-Bjerregaard, M. *et al.* A critical assessment of the microorganisms proposed to be important to enhanced biological phosphorus removal in full-scale wastewater treatment systems. *Front Microbiol* **8**, 718 (2017).
32 Hillebrand, H. On the generality of the latitudinal diversity gradient. *Am Nat* **163**, 192-211 (2004).
33 Martiny, J. B. H. *et al.* Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* **4**, 102-112 (2006).
34 Fuhrman, J. A. *et al.* A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci USA* **105**, 7774-7778, doi:10.1073/pnas.0803070105 (2008).
35 Zhou, J. *et al.* Temperature mediates continental-scale diversity of microbes in forest soils. *Nat Commun* **7**, 12083, doi:10.1038/ncomms12083 (2016).
36 Brown, J. H., Gillooly, J. F., Allen, A. P., Savage, V. M. & West, G. B. Toward a metabolic theory of ecology. *Ecology* **85**, 1771-1789 (2004).
37 Martiny, J. B., Eisen, J. A., Penn, K., Allison, S. D. & Horner-Devine, M. C. Drivers of bacterial beta-diversity depend on spatial scale. *Proc Natl Acad Sci USA* **108**, 7850-7854, doi:10.1073/pnas.1016308108 (2011).
38 Zhou, J. *et al.* Stochastic assembly leads to alternative communities with distinct functions in a bioreactor microbial community. *mBio* **4**, e00584-00512 (2013).
39 Knights, D. *et al.* Bayesian community-wide culture-independent microbial source tracking. *Nat Methods* **8**, 761-763 (2011).
40 Zhou, J. & Ning, D. Stochastic Community Assembly: Does It Matter in Microbial Ecology? *Microbiol Mol Biol Rev* **81**, e00002-00017 (2017).
41 Zhou, J. *et al.* Stochasticity, succession, and environmental perturbations in a fluidic ecosystem. *Proc Natl Acad Sci USA* **111**, E836-E845 (2014).
42 Hooper, D. U. *et al.* A global synthesis reveals biodiversity loss as a major driver of 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 Acevedo[1], Miriam Agullo-Barcelo[2], Gary L. Andersen[3,4], Juliana Calabria de Araujo[5], Kevin Boehnke[6], Philip Bond[2], Charles B. Bott[7], Patricia Bovio[8], Rebecca K. Brewster[6], Faizal Bux[9], Angela Cabezas[8], Léa Cabrol[10,11], Si Chen[12], Ting Chen[13,14], Claudia Etchebehere[8], Amanda Ford[7], Dominic Frigon[15], Janeth Sanabria Gómez[1], James S. Griffin[16], April Z. Gu[17], Moshe Habagil[18], Lauren Hale[19], Steven D. Hardeman[20], Marc Harmon[21], Harald Horn[22], Zhiqiang Hu[23], Shameem Jauffur[15,24], David R. Johnson[25], Alexander Keucken[18,26], Sheena Kumari[9], Cintia Dutra Leal[5], Laura A. Lebrun[27], Jangho Lee[28], Minjoo Lee[28], Zarraz MP Lee[29], Mengyan Li[30], Xu Li[31], Yu Liu[32,29], Richard G. Luthy[33], Leda C. Mendonça-Hagler[34], Francisca Gleire Rodriguez de Menezes[35], Arthur J. Meyers[36], Amin Mohebbi[31,37], Adrian Oehmen[38], Andrew Palmer[39], Prathap Parameswaran[40], Joonhong Park[28], Deborah Patsch[25], Valeria Reginatto[41], Francis L. de los Reyes III[42], Adalberto Noyola Robles[43], Simona Rossetti[44],

Jatinder Sidhu[39], William T. Sloan[45], Kylie Smith[39], Oscarina Viana de Sousa[35], Kyle Stephens[46], Renmao Tian[19], Nicholas B. Tooker[17], Daniel De los Cobos Vasconcelos[43], Steve Wakelin[47], Bei Wang[48], Joseph E. Weaver[42], Stephanie West[22], Paul Wilmes[27], Sung-Geun Woo[33], Jer-Horng Wu[49], Liyou Wu[19], Chuanwu Xi[6], Meiying Xu[50], Tao Yan[51], Min Yang[52], Michelle Young[40], Haowei Yue[53], Qian Zhang[51], Wen Zhang[54], Yu Zhang[52], Hongde Zhou[48]

Steering Committee: Jizhong Zhou[19,53,55], Xianghua Wen[53], Thomas P. Curtis[56], Qiang He[12,57], Zhili He[58,59]; **Contact**: Jizhong Zhou (jzhou@ou.edu), Daliang Ning[19,60,53] (ningdaliang@ou.edu)

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

606 Technological University, Singapore; [33]Department of Civil and Environmental Engineering,
607 Stanford University, Stanford, CA, USA; [34]Plant Biotechnology Program, Federal University of
608 Rio de Janeiro, UFRJ, Brazil; [35]Federal University of Ceará, UFC, Brazil; [36]University of
609 Tennessee, Center for Environmental Biotechnology, Knoxville, TN, USA; [37]Department of
610 Civil Engineering, Construction Management and Environmental Engineering, Northern Arizona
611 University, Flagstaff, AZ, USA; [38]UCIBIO, REQUIMTE, Departamento de Química, Faculdade
612 de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal; [39]CSIRO Land and
613 Water, Ecosciences Precinct, QLD, Australia; [40]Biodesign Swette Center for Environmental
614 Biotechnology, Arizona State University, Tempe, AZ, USA; [41]Departamento de Química,
615 Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto –
616 FFCLRP, Ribeirão Preto, SP, Brazil; [42]Department of Civil, Construction, and Environmental
617 Engineering, North Carolina State University, NC, USA; [43]Grupo de Investigación en Procesos
618 Anaerobios, Instituto de Ingeniería, Universidad Nacional Autónoma de México, México, DF,
619 Mexico; [44]CNR-IRSA, National Research Council, Water Research Institute, Rome, Italy;
620 [45]Infrastructure and Environment Research Division, School of Engineering, University of
621 Glasgow, Glasgow, UK; [46]Tryon Creek & Columbia Blvd. Wastewater Treatment Plants, Bureau
622 of Environmental Services, City of Portland, OR, USA; [47]Scion Research, Christchurch, New
623 Zealand; [48]School of Engineering, University of Guelph, Guelph, Ontario, Canada; [49]Department
624 of Environmental Engineering, National Cheng Kung University, Tainan City, Taiwan, ROC;
625 [50]State Key Laboratory of Applied Microbiology Southern China, Guangdong Provincial Key
626 Laboratory of Microbial Culture Collection and Application, Guangdong Institute of
627 Microbiology, Guangzhou, Guangdong, China; [51]Department of Civil and Environmental
628 Engineering, University of Hawaii at Manoa, Honolulu, HI, USA; [52]State Key Laboratory of
629 Environmental Aquatic Chemistry, Research Center for Eco-Environmental Sciences, Chinese
630 Academy of Sciences, Beijing, China; [53]State Key Joint Laboratory of Environment Simulation
631 and Pollution Control, School of Environment, Tsinghua University, Beijing, China;
632 [54]Department of Civil Engineering, University of Arkansas, Fayetteville, AR, USA; [55]Earth and
633 Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; [56]School
634 of Engineering, Newcastle University, Newcastle upon Tyne, UK; [57]Institute for a Secure and
635 Sustainable Environment, The University of Tennessee, Knoxville, TN, USA; [58]Environmental
636 Microbiomics Research Center, School of Environmental Science and Engineering, Sun Yat-Sen
637 University, Guangzhou, China; [59]Guangdong Provincial Key Laboratory of Environmental
638 Pollution Control and Remediation Technology, Guangzhou, China; [60]Consolidated Core
639 Laboratory, University of Oklahoma, Norman, Oklahoma, USA.
640

641 **Competing interests**

642

643 The authors declare no competing financial interests.
644

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 samples[5], 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 43.6ºS to 64.8 ºN. **(iii)** Climate zones. Since climate could have significant impacts on microbial communities, the samples covered 17

669 different climate types (Supplementary Fig. 6). To distinguish independent effects of continents
670 versus climate zones, we increased sampling efforts for climate zones that were present in
671 multiple continents, such as Humid Subtropical Climate. (**iv**) Scales. The samples were
672 collected from very broad spatial scales: global (across 6 continents), regional (e.g., individual
673 continents or climate zones), and local (e.g., individual cities). Within some cities, multiple
674 WWTPs and multiple samples per WWTP were collected; (**v**) Wastewater treatment process
675 types. To address the relationship of structure to function for activated sludge, we sampled the
676 aerobic zone of conventional plug flow, oxidation ditch, sequential batch reactors,
677 anaerobic/anoxic/oxic (A^2O), and other activated sludge process types.
678

679 A unified protocol was used for sampling, sample preservation, metadata collection, DNA
680 extraction, sequencing, and sequence analysis, to minimize potential experimental variations[4,51-
681 53]. Detailed sampling and metadata collection methods and protocols are available at the
682 GWMC web site (http://gwmc.ou.edu/protocols/view/11).
683

684 Sampling was carried out in June to November 2014 in the Northern Hemisphere and December
685 2014 to April 2015 in the Southern Hemisphere. The sampling time was generally between
686 10:00 am to 2:00 pm, when the WWTPs were relatively stable under normal conditions.
687 Although we tried to collect the global samples in the same season, seasonal temporal turnover in
688 activated sludge communities could have had some effect on the community variations we
689 observed to some degree. Based on limited published work[54,55], such temporal variations should
690 be much smaller than the spatial variations at the global/continental scales. For example, a
691 previous study on 5-year temporal dynamics of activated sludge community showed no

692 significant seasonal succession[54]. It's also revealed that the activated sludge communities were
693 relatively stable across three months, with average Bray-Curtis distance 0.45 ± 0.10 (mean ± SD)
694 between samples[55]; this variation was smaller than our observed mean variations even at local
695 city level (0.54 ± 0.19) (Fig. 4a).
696
697 At local scale, we defined a city based on it having a large enough geographic scale, not on an
698 administrative division (see Supplementary Table 1 for defined cities). For each city, we usually
699 collected at least 12 samples, and had ≥12 samples/city in 77% cities, with <3 samples/city in
700 only 1% of cities. We also sampled at least 2 WWTPs in 72% of the cities. In each plant, we
701 collected at least 3 mixed liquor samples, generally from 3 different positions (the front, middle,
702 and end part) of the aerobic zone in each aeration tank. In a few cases (3.3% plants), where only
703 one sampling position was applicable, 3 samples were taken in sequence with at least 30-min
704 interval. Altogether, we collected 1,186 activated sludge samples from 269 WWTPs across 23
705 countries from global scale (e.g., across 6 continents), regional scale (e.g., individual continents),
706 to local scale (e.g., geographic sites or individual cities) (Fig. 1a).
707
708 At each sampling position, approximately 1 liter mixed liquor was sampled and well mixed, and
709 40 mL was transferred into a sterile tube. The mixed liquor samples were kept on ice (≤4ºC),
710 transported to laboratory within 24 hours, divided into aliquots, and then centrifuged at 4ºC,
711 15,000 *g* for 10 min to collect pellets. Sludge pellets were transported (if necessary) with dry ice
712 to the designated laboratories within 48 hours and preserved at -80ºC before DNA extraction.
713

714 Along with the sludge samples, associated metadata, conforming to the Genomic Standards
715 Consortium's MIxS and Environmental Ontology (ENVO) Standards[56,57], were provided by
716 plant managers and/or investigators (Supplementary Table 1; Supplementary Fig. 7). We
717 collected metadata (e.g., chemical properties, operation conditions, process type) from each plant
718 using a standard sampling data sheet, which ensured that the data from all plants was in the same
719 format. Raw metadata were processed as one metadata table (Supplementary Table 1) and
720 classified into three categories: geological variables, plant operation and monitoring variables,
721 and sample properties. The geological variables included latitude and longitude; ambient climate
722 variables such as climate type, mean annual temperature (MAT), and precipitation; and
723 population size and gross domestic product (GDP) for the city where the WWTP was located.
724

725 Climate type was determined by the Köppen-Geiger climate classification[58]. GDP and
726 population data were derived from the Brookings analysis of Global Metro Monitor[59]. Variables
727 related to plant design and operation include plant age, design capacity, actual flow rate, volume
728 of aeration tanks, hydraulic retention time (HRT) and solids retention time (SRT). The activated
729 sludge process type, aerator type, and coupling with N removal processes (nitrification and
730 denitrification) in the WWTP were also provided by the plant managers as possible. Plant
731 monitoring variables include influent and effluent biochemical oxygen demand (BOD) and
732 chemical oxygen demand (COD) representing organic carbon (C) level, total nitrogen and total
733 phosphorus representing nutrient level, ammonium N, as well as the food to microorganism (F/M)
734 ratio, indicating the average organic C loading to microorganisms. For sample properties, most
735 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 (NH_4-N):

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

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 previously[60] and the primers 515F (GTGCCAGCMGCCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) of the Earth Microbiome Project[61]. *In silico* primer coverage analysis using SILVA TestPrime 1.0[62] and SILVA dataset r123 showed that these primers cover 86.8% and 52.9% of all bacterial and archaeal sequences with 0 mismatches, respectively.

779 To mitigate quantitative problems associated with amplicon sequencing[52], the 16S rRNA gene
780 fragments were amplified from community DNAs (10 ng) with two-step PCR using lower
781 numbers of amplification cycles (10 and 20 cycles for the 1st and 2nd step, respectively). The
782 two-step PAS approach offers several advantages: lower amplification biases, better sequence-
783 read quality, higher effective sequence read numbers and length, and lower sequencing errors[60].
784 All samples were sequenced using the same MiSeq instrument at the Institute for Environmental
785 Genomics, OU. Generally, about 400 samples were combined for each round of MiSeq
786 sequencing. Since the numbers of sequence reads varied substantially from sample to sample,
787 most samples were sequenced more than once (e.g., 19% twice; 33%, three times; 43%, > 3
788 times) to meet the target number of about 30K sequencing reads per sample, as determined in our
789 previous analysis[63].
790
791 The numbers of sequences (reads) per sample ranged from 25,631 to 351,844 (Supplementary
792 Table 5), and a total of 96,148 OTUs were obtained. About 1.3% of these OTUs were from
793 archaea, which accounted for 0.13% of the total abundance. The choice of the PCR primer pair
794 506F/806R (that was also used in the EMP project) is very likely to have strongly influenced this
795 low archaeal abundance due to the much lower coverage of the primers of archaeal 16S rRNA
796 genes compared to the bacterial counterparts. Because of the low archaeal abundance, the term
797 "bacteria" is used for simplicity. Also, the terms microbiome and microbial (or bacterial)
798 community are used interchangeably.
799
800 Raw sequence data were processed as previously described[35], except for OTU generation by
801 UPARSE[64] at the 97% similarity threshold, resulting in 96,148 OTUs. We define operational

802 taxonomic units (OTUs) (based on 97% sequence similarity) for bacterial and archaeal
803 phylotypes. Although there is potential misconnection between OTUs and microbial species[65],
804 we use this popular definition for simplicity, and it also allows us to compare with previous
805 studies of other systems. The representative sequences were aligned using Clustal Omega
806 v1.2.2[66] for constructing the phylogenetic tree by FastTree2 v2.1.10[67]. OTUs were
807 taxonomically annotated by RDP Classifier[68] with a confidence cutoff of 80%, using the MiDAS
808 database (version 2.1) which specifically provides a curated taxonomy for abundant and
809 functionally important microorganisms in activated sludge[69]. After removal of the global
810 singletons[64], the sequence number in each sample was rarefied to the same depth (25,600
811 sequences per sample), resulting in 61,448 OTUs overall, which were used in subsequent
812 comparative analyses.
813
814 Although our sequencing depths were considerably higher than those in many similar studies[70],
815 rarefaction curves (Supplementary Fig. 2d, e) of activated sludge microbial communities
816 indicated that additional rare taxa were likely present in individual samples. Nevertheless,
817 pooling all sequences gave a sufficient number for estimating global- and continent-level
818 diversity of activated sludge microbial communities (Supplementary Fig. 2f, g). The global
819 OTU richness per sample was 2,309±559 (Supplementary Table 5). Besides richness, we also
820 calculated other alpha diversity indices on a global and regional scale (Supplementary Table 5).
821
822 The rRNA operon copy number for each OTU was estimated through the rrnDB database based
823 on its closest relatives with known rRNA operon copy number[71]. The abundance-weighted mean
824 rRNA operon copy number was then calculated for each sample as described previously[49].

825

826 **Sequence comparison against reference databases**

827

828 To compare the sequence diversity in this study to that in existing databases, the 96,148
829 representative sequences from the activated sludge samples were compared against the
830 representative set (97% similarity level) of full-length sequences from Greengenes 13.8[72]
831 (released on August 2013) and the non-eukaryotic fraction of Silva 132 databases[73] (released on
832 December 2017). We used the open-source sequence search tool USEARCH10[74] in global
833 alignment search mode, and we required 97% similarity across the query sequence. Our
834 activated sludge sequences match to 38.6% of Greengenes and 37.2% of SILVA 16S rRNA gene
835 OTUs at 97% similarity. These matches accounted for 18.2% and 22.5% of the representative
836 sequences in our datasets, respectively, indicating that the majority of activated sludge microbial
837 species diversity is not yet captured in full-length sequence databases; this is similar to the
838 observations in the EMP[5].

839

840 **Species abundance distribution (SAD) fitting**

841

842 We compared the SAD of each sample, based on the rank-abundance distribution, with
843 predictions from Poisson lognormal, log-series, Broken-stick, and Zipf models. Although
844 numerous SAD models are available, lognormal and log-series have been the most successful in
845 predicting SADs, and they are the standards for testing other models[18]. While the logseries
846 model is well supported by macroecological studies, the Poisson lognormal model is more
847 commonly observed with microorganisms[18]. 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 al[18].

Estimation of global bacterial diversity of WWTPs

We used the methods described in Curtis et al.[20] and Locey and Lennon[19] 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:

$$S_T = \frac{\sqrt{\pi}}{a} exp\left\{\left(alog_2\left(\sqrt{\frac{N_{max}}{N_{min}}}\right)\right)^2\right\} \quad \text{(i)}$$

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

$$N_T = \frac{\sqrt{\pi N_{min} N_{max}}}{2a} exp\left\{\left(alog_2\left(\sqrt{\frac{N_{max}}{N_{min}}}\right)\right)^2\right\} exp\left\{\left(\frac{ln(2)}{2a}\right)^2\right\}\left[erf\left(alog_2\left(\sqrt{\frac{N_{max}}{N_{min}}} - \frac{ln(2)}{2a}\right)\right) + erf\left(alog_2\left(\sqrt{\frac{N_{max}}{N_{min}}} + \frac{ln(2)}{2a}\right)\right)\right] \quad \text{(ii)}$$

We used published data to estimate the total microbial abundance in WWTPs as follows. Empirical records compiled from a variety of sources, for example, AQUASTAT[75] and Sato et al 2013[76], suggest that about 330 km^3 $year^{-1}$ of municipal wastewater are produced globally, of

869 which 60% is treated[9]. Assuming that they are all treated in WWTPs, then about 0.54 km^3
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 $V = Q \times HRT$ (iii)
873 where Q is the influent flow rate (m^3 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) km^3. The total cells in activated
876 sludge are about 2.3 (± 0.4)× 10^9 (ml^{-1})[77]; thus, N_T (global activated sludge bacterial abundance)
877 is about 4.0- 6.1 × 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 law[19]. 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 m^3, making the total cells about 2.3 (± 0.4) × 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 law[19]. The total bacterial richness estimates of individual
891 human gut, individual cow rumen, global ocean and Earth were taken from Locey and Lennon[19].

892

893 **Core community determination**

894 A global-scale core microbial community was determined based on multiple reported measures.
895 First, "overall abundant OTUs" were filtered out according to mean relative abundance across all
896 samples (MRA)[80]. Previous studies used different criteria (e.g., MRA > 1%[30,81] or 0.1%[82,83])
897 without any objective or standard rule. Thus, we selected all top 0.1% OTUs (62) as overall
898 abundant OTUs. Their MRA was higher than 0.2%, within the range of reported criteria.
899 Second, "ubiquitous OTUs" were defined as OTUs with occurrence frequency in more than 80%
900 of all samples[84]. Finally, "frequently abundant OTUs" were selected based on their relative
901 abundances with a sample. In each sample, the OTUs were defined as abundant when they had a
902 higher relative abundance than other OTUs and made up the top 80% of the reads in the sample[14].
903 A frequently abundant OTU was defined as abundant in at least half samples, which is stricter
904 than the reported criterion (10 in 26 samples[14]). Since the above three measures are
905 complementary to one another when defining core community, only OTUs fulfilling all three
906 criteria were defined as the global scale core bacterial community.
907
908 Following the same criteria as described above, the core community was identified for each
909 continent. That is, a core OTU for a specific continent should be one that was from the top 0.1%
910 OTUs of that continent; a core OTU also had to be detected in more than 80% of the samples and
911 dominant for more 50% of the samples of that continent.
912

913 **Comparison of bacterial community composition of WWTPs to natural habitats and source**
914 **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_greengenes/emp_cr_gg_13_8.subset_
918 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 SourceTracker[39], 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 package[85]. Other taxonomic α-diversity indices, including Shannon index, Simpson index and Pielou's evenness, were calculated using the *vegan* R package[86].

Bray-Curtis (abundance-based) and Sorensen (incidence-based) distances were calculated to represent the taxonomic β-diversity using the *vegan* R package[86]. Canberra's distance was also calculated to give more weight to rare taxa, using the *vegan* R package[86]. The weighted (abundance-based) and unweighted UniFrac (incidence-based) distance[87] were calculated to represent the phylogenetic β-diversity usig the *GUniFrac* R package[88]. 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 package[86].

PERMANOVA was applied to assess the difference of community composition among continents, climate types, and activated sludge process types using the *vegan* R package[86]. 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 (A^2O), anoxic/oxic (AO), oxidation ditch, contact stabilization, pure oxygen and extended aeration.

961

962 **Distance Decay relationships**

963

964 The rate of the distance-decay relationship (DDR) was calculated as the slope of a linear least
965 squares regression on the relationship between ln-transformed geographic distance versus ln-
966 transformed bacterial community composition similarity. We used matrix permutation tests to
967 examine the statistical significance of the distance-decay slope[37]. The samples were permuted
968 999 times, and the observed slope was compared with the distribution of values in the permuted
969 datasets. We also tested whether the slopes of the distance-decay curve at the three spatial scales
970 (0 to 100 km; 100 to 5,000 km; and 5,000 to 25,000 km) were significantly different from the
971 slope of the overall distance-decay curve, using matrix permutations to compare the observed
972 difference between slopes within the three spatial scales with the overall distance-decay slope to
973 that over 999 permutations.

974

975 **Estimating stochasticity of community assembly**

976

977 We assessed community-assembly stochasticity with a null-model-based index. The
978 Stochasticity ratio was described previously[41,89]. Since null-model algorithms usually require a
979 high number of replicates, we selected 71 cities, each of which had more than 9 samples; we
980 randomly drew 9 samples from each city to make sampling even. We calculated stochasticity
981 ratio using taxonomic and phylogenetic metrics. Whether using the Bray-Curtis (abundance-
982 weighted) or Sorensen (unweighted) model, the stochasticity ratio was calculated based on
983 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 previously[37]. 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 test[37]. The significance of the partial regression was tested by matrix permutation for 999 times[93]. 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 elsewehere[94]: (i) pure environmental variation = $R^2_T - R^2_G$; (ii) pure geographical distance = $R^2_T - R^2_E$; (iii)

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

Structural equation model (SEM)

SEM was used to explore the direct and indirect relationships among environmental variables, bacterial communities, and activated sludge function. The community composition was represented by the first principal component (PC1) of Principal coordinate analysis (PCoA) based on Bray-Curtis distance. We first considered a full model that included all reasonable pathways, and then we sequentially eliminated non-significant pathways until we attained the final model whose pathways all were significant. To capture the quadratic correlation of SRT to diversity and BOD removal, we constructed a composite variable[94] of 'SRT effect' as a linear combination of SRT and the square of SRT (SRT.SQ). We used a χ^2 test and the root mean square error of approximation to evaluate the fit of model. The SEM-related analysis was performed using the *lavaan* R package[95].

Random Forest models

We applied a machine-learning model, random forest, to examine the strengths of the associations between environmental variable and compositional data, using the *randomForest* R package[96]. We used OTUs as predictors and environmental variable as response data. To correct the potential spatial autocorrelation, we used OTU data at the plant level, by averaging 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).

53 Sinha, R. *et al.* Assessment of variation in microbial community amplicon sequencing by the Microbiome Quality Control (MBQC) project consortium. *Nat Biotechnol* **35**, 1077 (2017).
54 Ju, F. & Zhang, T. Bacterial assembly and temporal dynamics in activated sludge of a full-scale municipal wastewater treatment plant. *ISME J* **9**, 683-695 (2015).
55 Xia, Yu. *Diversity and temporal assembly patterns of microbial communities in municipal wastewater treatment systems*. PhD thesis, Univ. Tsinghua, Beijing, China (2016).
56 Buttigieg, P. L., Morrison, N., Smith, B., Mungall, C. J. & Lewis, S. E. The environment ontology: contextualising biological and biomedical entities. *J Biomed Semantics* **4**, 43 (2013).
57 Yilmaz, P. *et al.* Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol* **29**, 415-420 (2011).
58 Peel, M. C., Finlayson, B. L. & McMahon, T. A. Updated world map of the Köppen-Geiger climate classification. *Hydrol Earth Syst Sci Discuss* **4**, 439-473 (2007).
59 Alan Berube, J. L. T., Tao Ran,Joseph Parilla. *Global Metro Monitor*, <https://www.brookings.edu/research/global-metro-monitor/> (2015).
60 Wu, L. *et al.* Phasing amplicon sequencing on Illumina Miseq for robust environmental microbial community analysis. *BMC Microbiol* **15**, 125 (2015).
61 Caporaso, J. G. *et al.* Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proc Natl Acad Sci USA* **108**, 4516-4522 (2011).
62 Klindworth, A. *et al.* Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res* **41**, e1-e1 (2013).
63 Wen, C. *et al.* Evaluation of the reproducibility of amplicon sequencing with Illumina MiSeq platform. *PLoS One* **12**, e0176716 (2017).
64 Edgar, R. C. UPARSE: highly accurate OTU sequences from microbial amplicon reads. *Nat Methods* **10**, 996-998 (2013).
65 McLaren, M. R. & Callahan, B. J. In Nature, There Is Only Diversity. *mBio* **9**, e02149-02117 (2018).
66 Sievers, F. *et al.* Fast, scalable generation of high□quality protein multiple sequence alignments using Clustal Omega. *Mol Syst Biol* **7**, doi:10.1038/msb.2011.75 (2011).
67 Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2 – Approximately maximum-likelihood trees for large alignments. *PLoS One* **5**, e9490, doi:10.1371/journal.pone.0009490 (2010).
68 Wang, Q., Garrity, G. M., Tiedje, J. M. & Cole, J. R. Naive Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. *Appl Environ Microbiol* **73**, 5261-5267 (2007).
69 McIlroy, S. J. *et al.* MiDAS 2.0: an ecosystem-specific taxonomy and online database for the organisms of wastewater treatment systems expanded for anaerobic digester groups. *Database* **2017** (2017).
70 Delgado-Baquerizo, M. *et al.* A global atlas of the dominant bacteria found in soil. *Science* **359**, 320-325 (2018).

71 Stoddard, S. F., Smith, B. J., Hein, R., Roller, B. R. & Schmidt, T. M. rrnDB: improved tools for interpreting rRNA gene abundance in bacteria and archaea and a new foundation for future development. *Nucleic Acids Res*, gku1201 (2014).
72 McDonald, D. *et al.* An improved Greengenes taxonomy with explicit ranks for ecological and evolutionary analyses of bacteria and archaea. *ISME J* **6**, 610 (2012).
73 Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res* **41**, D590-D596 (2012).
74 Edgar, R. C. Search and clustering orders of magnitude faster than BLAST. *Bioinformatics* **26**, 2460-2461 (2010).
75 AQUASTAT. *FAO global information system on water and agriculture. Wastewater section.*, <http://www.fao.org/nr/water/aquastat/wastewater/index.stm> (2014).
76 Sato, T., Qadir, M., Yamamoto, S., Endo, T. & Zahoor, A. Global, regional, and country level need for data on wastewater generation, treatment, and use. *Agri Water Manag* **130**, 1-13 (2013).
77 Foladori, P., Bruni, L., Tamburini, S. & Ziglio, G. Direct quantification of bacterial biomass in influent, effluent and activated sludge of wastewater treatment plants by using flow cytometry. *Water Res* **44**, 3807-3818 (2010).
78 Agency, U. S. E. P. *The Sources and Solutions: Wastewater*, <https://www.epa.gov/nutrientpollution/sources-and-solutions-wastewater> (2018).
79 Chan, W. *Wastewater: Good To The Last Drop*, <http://chinawaterrisk.org/resources/analysis-reviews/wastewater-good-to-the-last-drop/> (2017).
80 Hanski, I. Dynamics of regional distribution: the core and satellite species hypothesis. *Oikos*, 210-221 (1982).
81 Galand, P. E., Casamayor, E. O., Kirchman, D. L. & Lovejoy, C. Ecology of the rare microbial biosphere of the Arctic Ocean. *Proc Natl Acad Sci USA* **106**, 22427-22432, doi:10.1073/pnas.0908284106 (2009).
82 Székely, A. J. & Langenheder, S. The importance of species sorting differs between habitat generalists and specialists in bacterial communities. *FEMS Microbiol Ecol* **87**, 102-112 (2014).
83 Cheng, J. *et al.* Discordant temporal development of bacterial phyla and the emergence of core in the fecal microbiota of young children. *ISME J* **10**, 1002 (2016).
84 Ju, F. & Zhang, T. Bacterial assembly and temporal dynamics in activated sludge of a full-scale municipal wastewater treatment plant. *ISME J*, doi:10.1038/ismej.2014.162 (2014).
85 Kembel, S. W. *et al.* Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463-1464 (2010).
86 Oksanen, J. *et al.* Package 'vegan'. *Community ecology package, version* **2** (2013).
87 Lozupone, C. & Knight, R. UniFrac: a new phylogenetic method for comparing microbial communities. *Appl Environ Microbiol* **71**, 8228-8235 (2005).
88 Chen, J. GUniFrac: generalized UniFrac distances. *R package version* **1**, 2012 (2012).
89 Guo, X. *et al.* Climate warming leads to divergent succession of grassland microbial communities. *Nat Clim Change* **8**, 813 (2018).
90 Chase, J. M., Kraft, N. J., Smith, K. G., Vellend, M. & Inouye, B. D. Using null models to disentangle variation in community dissimilarity from variation in α-diversity. *Ecosphere* **2**, art24 (2011).

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

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 Lennon[19]. 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 law[19] (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. $P < 2\times10^{-16}$ (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 (NH_4-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

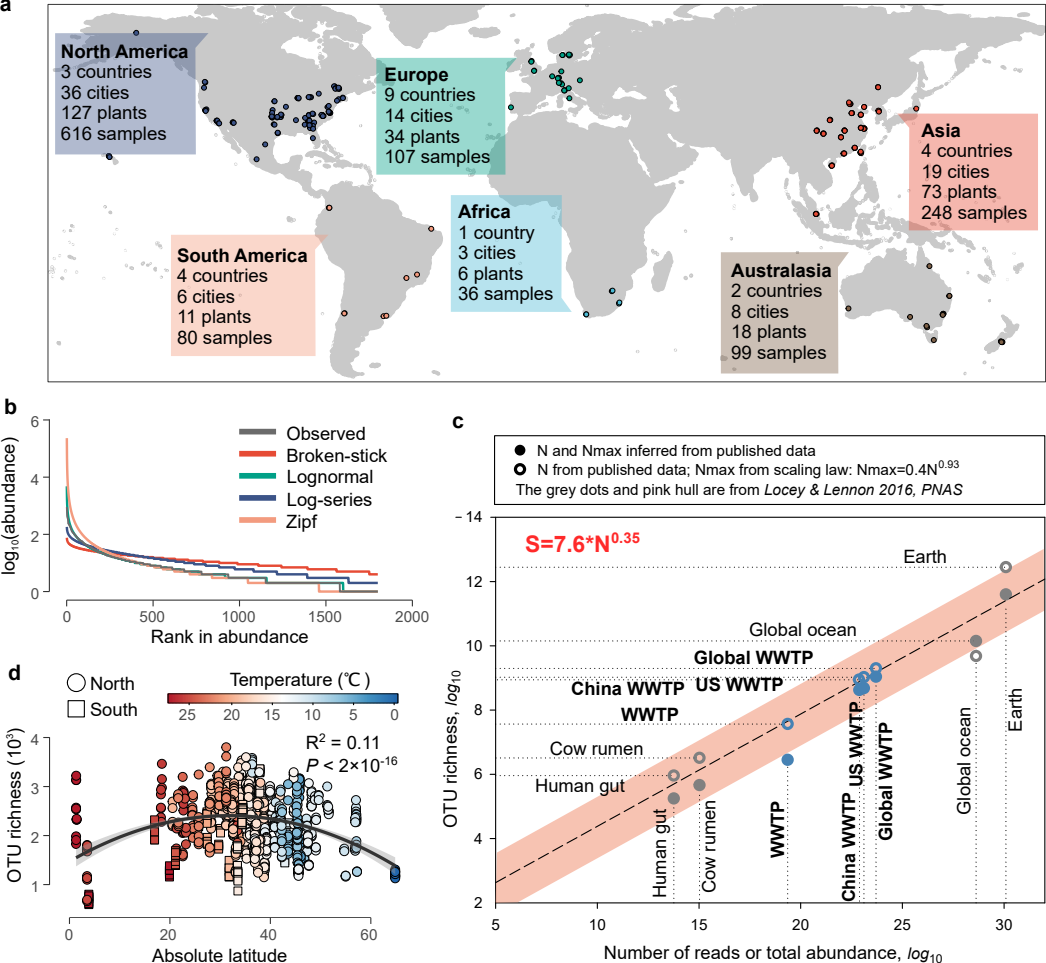
discovery rate (FDR) controlling procedure (n = 14,235 pairwise cases). Only global core OTUs were shown, with their mean relative abundance indicated on the left of the heatmap.

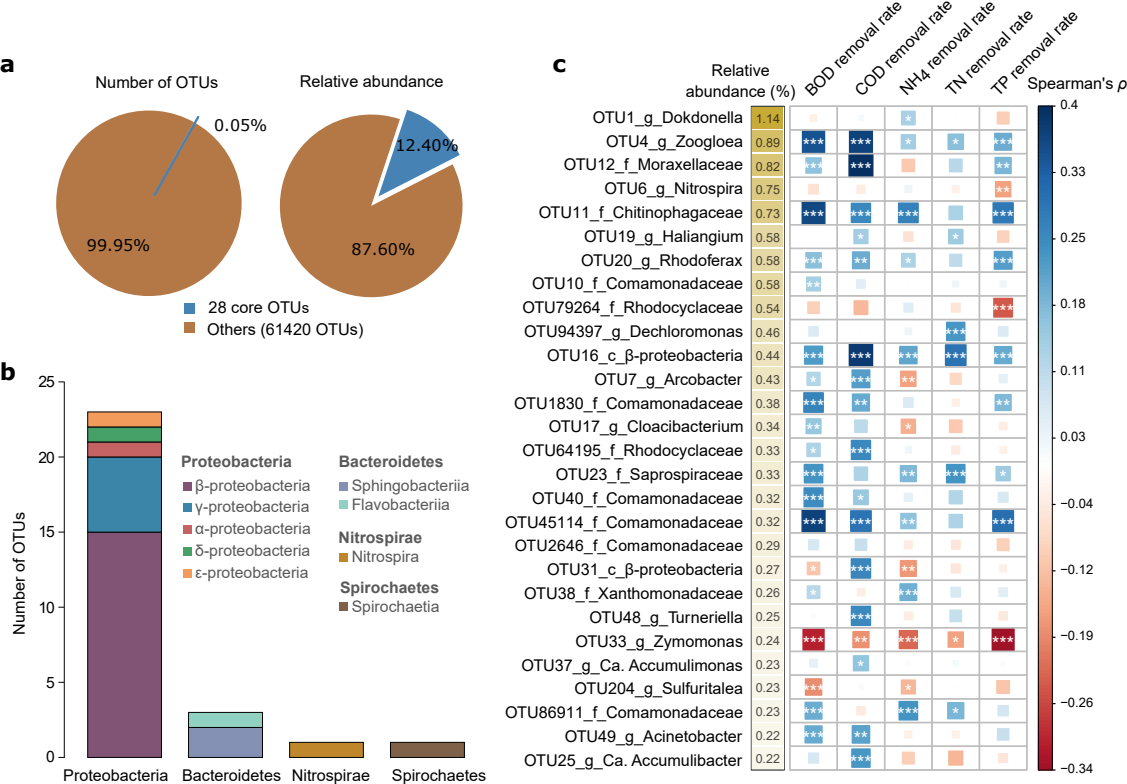
Fig. 3. Comparing bacterial community compositions across continents and with other habitats. (**a**) Nonmetric Multidimensional Scaling analysis (NMDS) showing that activated sludge of WWTPs harbored a unique microbiome as compared with other habitats. For comparison, we merged our OTU table (n = 269 WWTPs) with that released by EMP[5], 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 * 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 * 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.

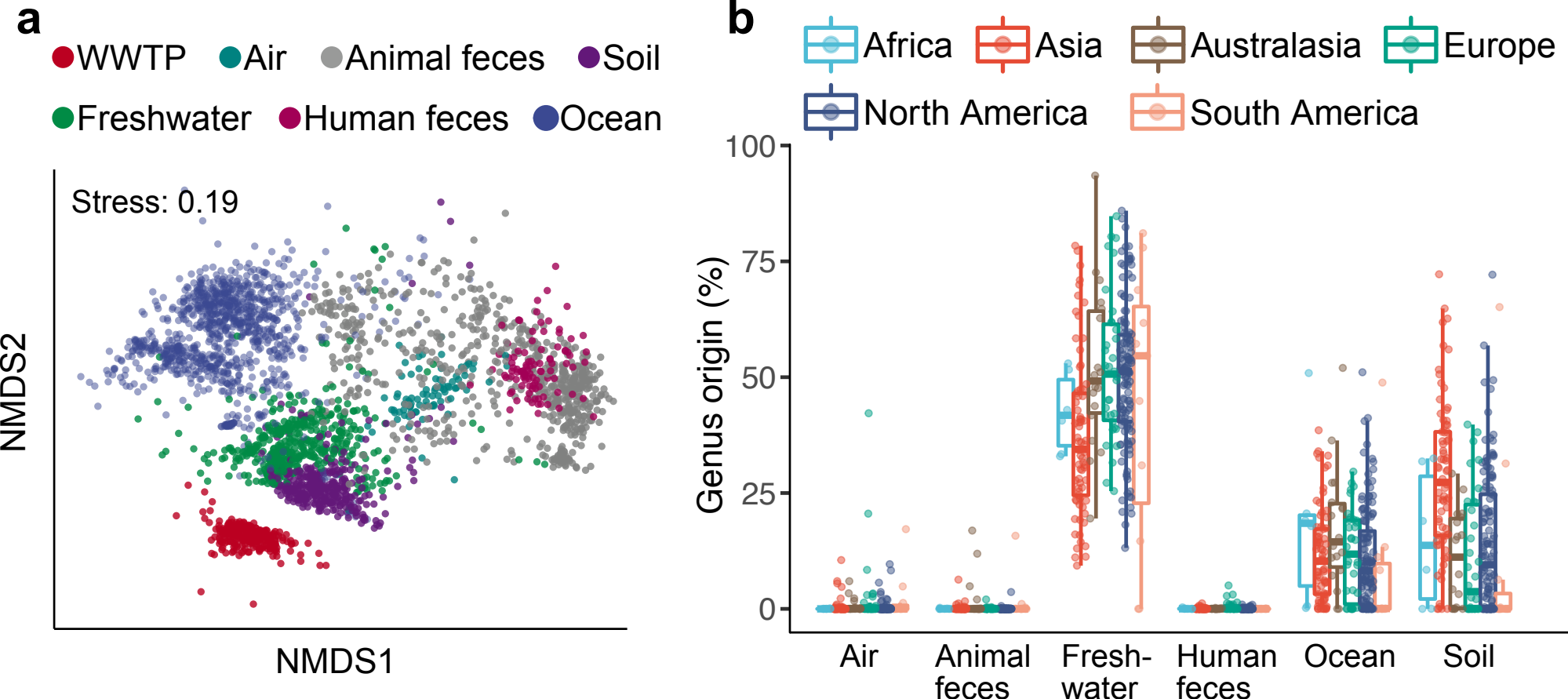
Fig. 4. Spatial turnover of the activated sludge bacterial communities. (a) Distance-decay relationships (DDRs) based on Bray-Curtis similarity. Black line denotes the least-squares linear regression across all spatial scales (n= 702,705 pairwise distances). Colored lines denote separate regressions: 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.

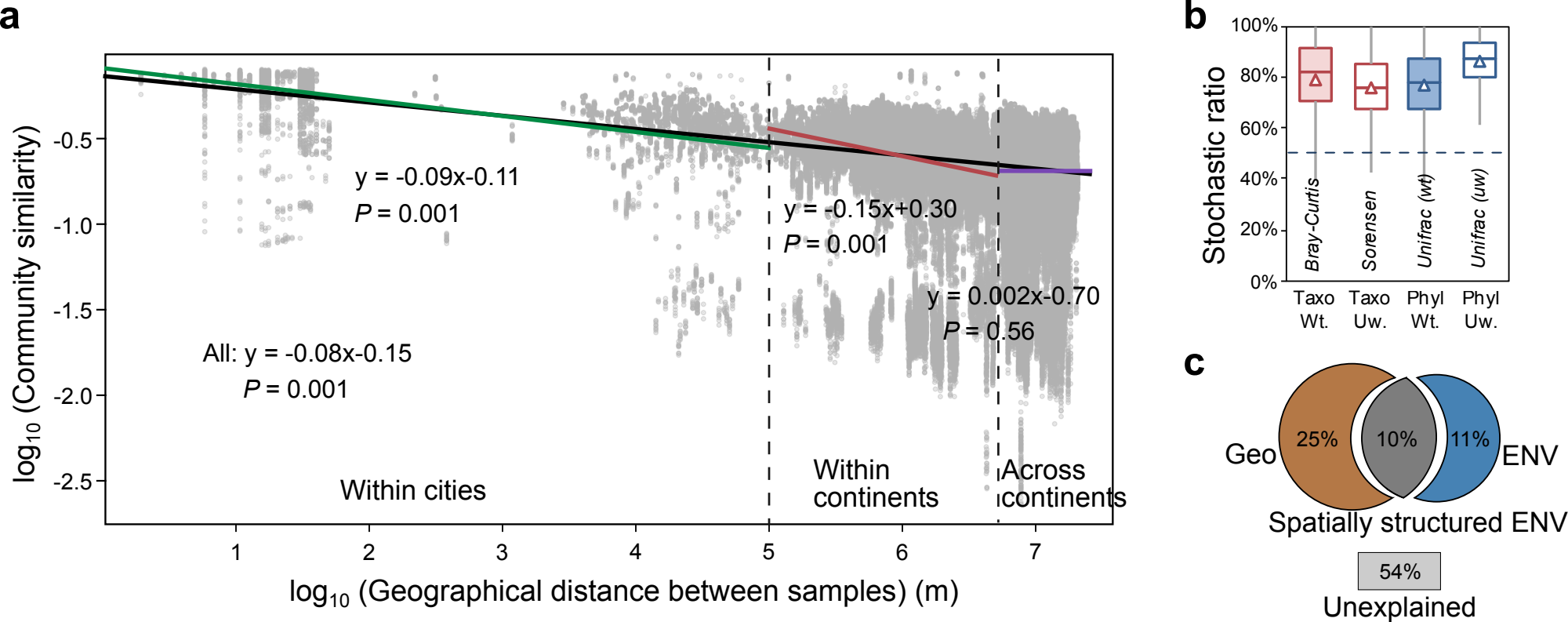
Fig. 5. Environmental drivers of the activated sludge community composition. (**a**) A structural equation model (SEM) shows relationships among environmental variables, community composition, and WWTP functioning. The composite variable of 'SRT effect' was constructed as a linear combination of solids 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

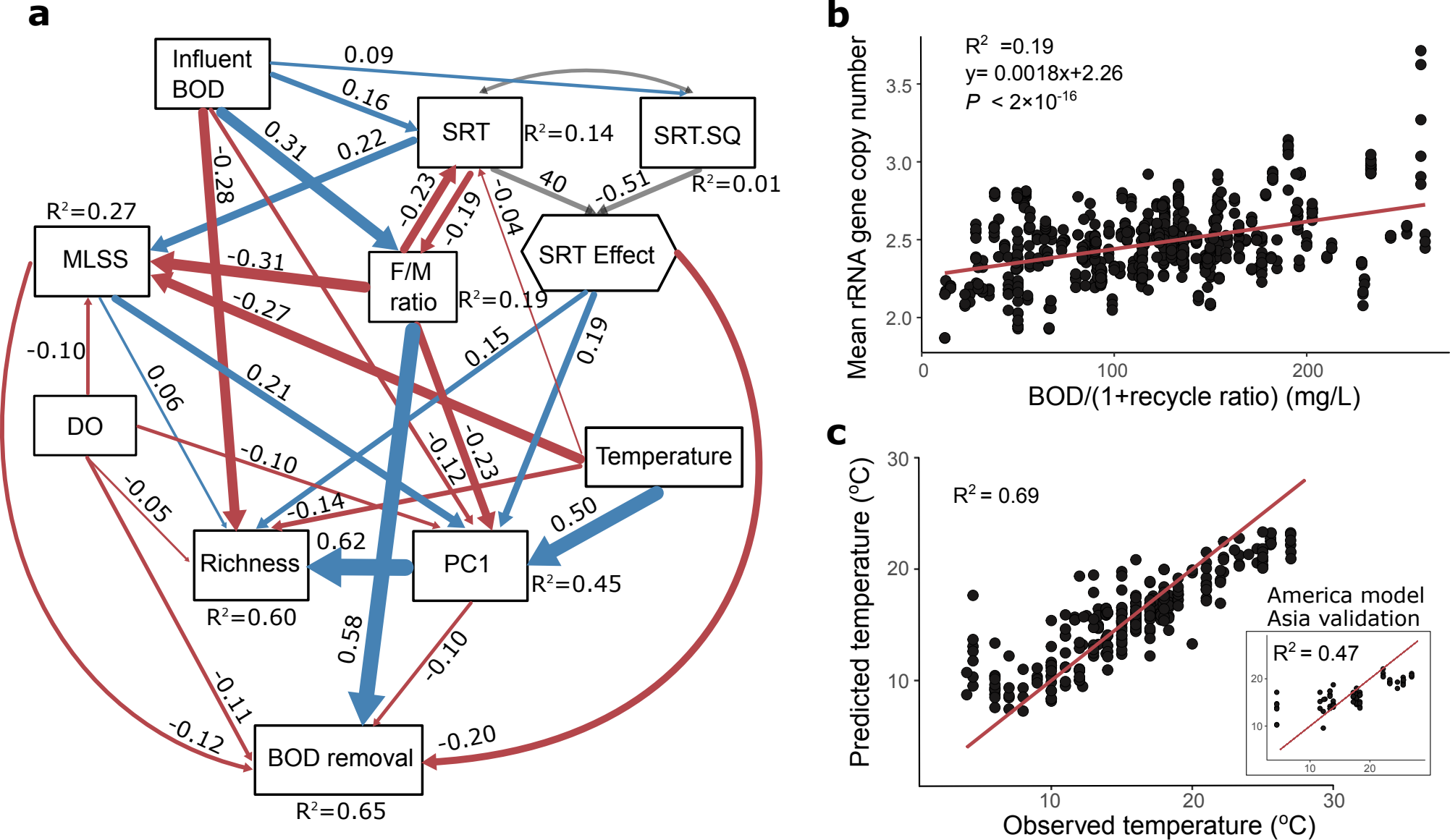
(β). Arrow width is proportional to the strength of the relationship. R^2 represents the proportion of variance explained for every dependent variable. Model $\chi^2 = 13.92$, df = 12, $P = 0.31$, n = 1,186 biologically independent samples; root mean square error of approximation (RMSEA) = 0.012 with 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
North America
3 countries
36 cities
127 plants
616 samples
Europe
9 countries
14 cities
34 plants
107 samples
Asia
4 countries
19 cities
73 plants
248 samples
Africa
1 country
3 cities
6 plants
36 samples
South America
4 countries
6 cities
11 plants
80 samples
Australasia
2 countries
8 cities
18 plants
99 samples
b
Observed
Broken-stick
Lognormal
Log-series
Zipf
log10(abundance)
Rank in abundance
c
N and Nmax inferred from published data
N from published data; Nmax from scaling law: Nmax=0.4N^0.93
The grey dots and pink hull are from Locey & Lennon 2016, PNAS
S=7.6*N^0.35
Earth
Global ocean
Global WWTP
US WWTP
China WWTP
WWTP
Cow rumen
Human gut
OTU richness, log10
Number of reads or total abundance, log10
d
North
South
Temperature (℃)
R2 = 0.11
P < 2×10-16
OTU richness (10^3)
Absolute latitude


a
Number of OTUs
0.05%
99.95%
Relative abundance
12.40%
87.60%
28 core OTUs
Others (61420 OTUs)
b
Number of OTUs
25
20
15
10
5
0
Proteobacteria
Bacteroidetes
Nitrospirae
Spirochaetes
Proteobacteria
β-proteobacteria
γ-proteobacteria
α-proteobacteria
δ-proteobacteria
ε-proteobacteria
Bacteroidetes
Sphingobacteriia
Flavobacteriia
Nitrospirae
Nitrospira
Spirochaetes
Spirochaetia
c
Relative abundance (%)
BOD removal rate
COD removal rate
NH4 removal rate
TN removal rate
TP removal rate
Spearman's ρ
0.4
0.33
0.25
0.18
0.11
0.03
−0.04
−0.12
−0.19
−0.26
−0.34
OTU1_g_Dokdonella 1.14
OTU4_g_Zoogloea 0.89
OTU12_f_Moraxellaceae 0.82
OTU6_g_Nitrospira 0.75
OTU11_f_Chitinophagaceae 0.73
OTU19_g_Haliangium 0.58
OTU20_g_Rhodoferax 0.58
OTU10_f_Comamonadaceae 0.58
OTU79264_f_Rhodocyclaceae 0.54
OTU94397_g_Dechloromonas 0.46
OTU16_c_β-proteobacteria 0.44
OTU7_g_Arcobacter 0.43
OTU1830_f_Comamonadaceae 0.38
OTU17_g_Cloacibacterium 0.34
OTU64195_f_Rhodocyclaceae 0.33
OTU23_f_Saprospiraceae 0.33
OTU40_f_Comamonadaceae 0.32
OTU45114_f_Comamonadaceae 0.32
OTU2646_f_Comamonadaceae 0.29
OTU31_c_β-proteobacteria 0.27
OTU38_f_Xanthomonadaceae 0.26
OTU48_g_Turneriella 0.25
OTU33_g_Zymomonas 0.24
OTU37_g_Ca. Accumulimonas 0.23
OTU204_g_Sulfuritalea 0.23
OTU86911_f_Comamonadaceae 0.23
OTU49_g_Acinetobacter 0.22
OTU25_g_Ca. Accumulibacter 0.22


a
WWTP
Air
Animal feces
Soil
Freshwater
Human feces
Ocean
Stress: 0.19
NMDS2
NMDS1
b
Africa
Asia
Australasia
Europe
North America
South America
Genus origin (%)
100
75
50
25
0
Air
Animal feces
Fresh-water
Human feces
Ocean
Soil


a
log10 (Community similarity)
-0.5
-1.0
-1.5
-2.0
-2.5
y = -0.09x-0.11
P = 0.001
y = -0.15x+0.30
P = 0.001
y = 0.002x-0.70
P = 0.56
All: y = -0.08x-0.15
P = 0.001
Within cities
Within continents
Across continents
1
2
3
4
5
6
7
log10 (Geographical distance between samples) (m)
b
Stochastic ratio
100%
80%
60%
40%
20%
0%
Bray-Curtis
Sorensen
Unifrac (wt)
Unifrac (uw)
Taxo Wt.
Taxo Uw.
Phyl Wt.
Phyl Uw.
c
Geo
25%
10%
11%
ENV
Spatially structured ENV
54%
Unexplained


a
Influent BOD
SRT
R²=0.14
SRT.SQ
R²=0.01
SRT Effect
MLSS
R²=0.27
F/M ratio
R²=0.19
DO
Temperature
Richness
R²=0.60
PC1
R²=0.45
BOD removal
R²=0.65
0.09
0.16
0.31
0.22
-0.28
-0.23
-0.19
-0.04
40
-0.51
-0.31
-0.27
-0.10
0.06
0.21
0.15
0.19
-0.12
-0.23
-0.10
-0.05
-0.14
0.62
0.50
0.58
-0.10
-0.11
-0.12
-0.20
b
R² =0.19
y= 0.0018x+2.26
P < 2×10-16
Mean rRNA gene copy number
BOD/(1+recycle ratio) (mg/L)
c
R² = 0.69
Predicted temperature (°C)
Observed temperature (°C)
America model
Asia validation
R² = 0.47