nature microbiology

ARTICLES

https://doi.org/10.1038/s41564-019-0426-5

Global diversity and biogeography of bacterial communities in wastewater treatment plants

Linwei Wu[1,2,29], Daliang Ning[1,2,3,29], Bing Zhang[1,2,29], 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[14], Tong Zhang[15], Zhili He[16,17], Jurg Keller[18], Per H. Nielsen[19], Pedro J. J. Alvarez[20], Craig S. Criddle[21], Michael Wagner[9], James M. Tiedje[22], Qiang He[23,24]*, Thomas P. Curtis[6]*, David A. Stahl[25], Lisa Alvarez-Cohen[26,27], Bruce E. Rittmann[28], Xianghua Wen[1]* and Jizhong Zhou[1,2,27]*

Microorganisms in wastewater treatment plants (WWTPs) are essential for water purification to protect public and environmental health. However, the diversity of microorganisms and the factors that control it are poorly understood. Using a systematic global-sampling effort, we analysed the 16S ribosomal RNA 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 ($n = 28$ operational taxonomic units) 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 and drift), although deterministic factors (temperature and organic input) are also important. Our findings enhance our 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.

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. Unravelling the mechanisms that generate and underlie 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, fresh water, air and soil.

[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, OK, USA. [4]College of Resource and 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, 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 and Interuniversity Cooperation Centre for Water and Health, Krems, Austria. [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]A full list of Global Water Microbiome Consortium members appears at the end of the paper. [15]Environmental Biotechnology Laboratory, The University of Hong Kong, Hong Kong, China. [16]Environmental Microbiomics Research Center, School of Environmental Science and Engineering, Sun Yat-Sen University, Guangzhou, China. [17]Guangdong Provincial Key Laboratory of Environmental Pollution Control and Remediation Technology, Guangzhou, China. [18]Advanced Water Management Centre, The University of Queensland, Brisbane, Queensland, Australia. [19]Department of Chemistry and Bioscience, Center for Microbial Communities, Aalborg University, Aalborg, Denmark. [20]Department of Civil and Environmental Engineering, Rice University, Houston, TX, USA. [21]Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA. [22]Center for Microbial Ecology, Michigan State University, East Lansing, MI, USA. [23]Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN, USA. [24]Institute for a Secure and Sustainable Environment, The University of Tennessee, Knoxville, TN, USA. [25]Department of Civil and Environmental Engineering, University of Washington, Seattle, WA, USA. [26]Department of Civil and Environmental Engineering, College of Engineering, University of California, Berkeley, CA, USA. [27]Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [28]Biodesign Swette Center for Environmental Biotechnology, Arizona State University, Tempe, AZ, USA. [29]These authors contributed equally: Linwei Wu, Daliang Ning, Bing Zhang. *e-mail: qianghe@utk.edu; tom.curtis@newcastle.ac.uk; xhwen@tsinghua.edu.cn; jzhou@ou.edu

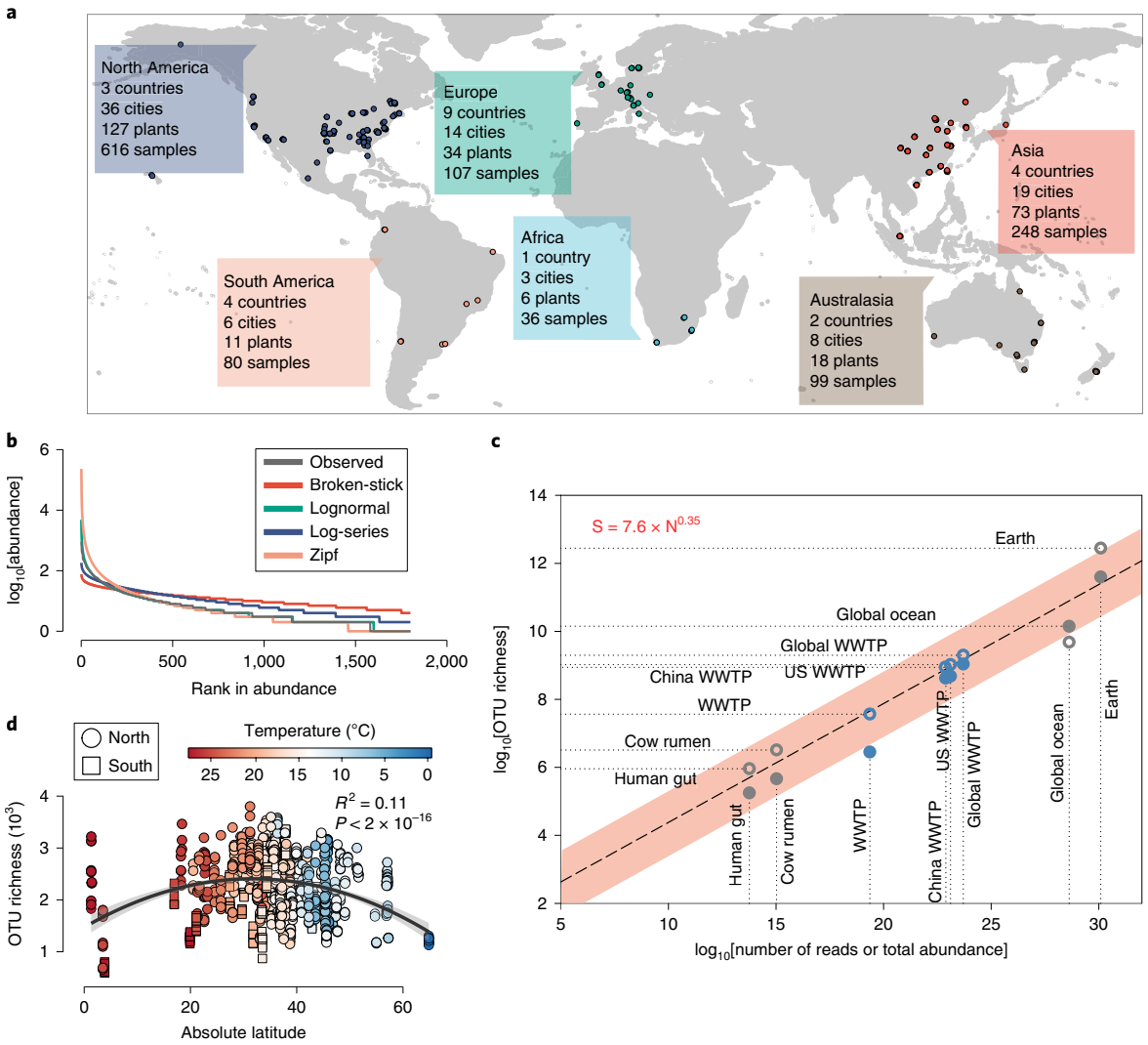

Fig. 1 | The GWMC captures microbial diversity of globally distributed WWTPs. a, Geographical distribution of 269 WWTPs where activated sludge samples and environmental data were collected. **b**, Predicting SAD of activated sludge bacterial communities. The grey line represents a SAD that was randomly chosen from our data. Each model was fit to the observed SAD (see Methods). Supplementary Table 2 shows the variations of the SADs explained by each model across all 1,186 activated sludge communities, indicating the best performance of the Poisson lognormal model. **c**, Estimation of the microbial richness of activated sludge of WWTPs. Microbial species are defined as OTUs at 97% sequence similarity threshold. The microbial richness (S)–abundance (N) scaling relationship (broken grey line with pink shading as 95% prediction interval), and the grey circles representing richness estimates from other systems were derived from a previous study[19]. Richness was predicted from the lognormal model using N_T estimated from published data, and N_{max} inferred from our sequencing data (filled circles) or $N_{max} = 0.4 \times N_T^{0.93}$ predicted from the dominance-scaling law[19] (open circles). WWTP indicates one WWTP, as do Human gut and Cow rumen. **d**, Latitudinal distribution of activated sludge bacterial diversity, plotting OTU richness against the absolute latitude of sampling locations shows the peak of richness at intermediate latitude ($n = 1{,}186$ biologically independent samples). The line shows the second-order polynomial fit based on ordinary least squares regression. $P < 2 \times 10^{-16}$ (two-sided) for both regression coefficients. The colour gradient denotes the annual mean air temperature. Shapes of symbols denotes whether a sample originated from the Northern Hemisphere or the Southern Hemisphere.

However, we are just beginning to understand the diversity and biogeography of microbial communities in wastewater treatment plants (WWTPs)[3,8].

More than 300 km³ of wastewater is produced globally each year[9]. This volume equals one-seventh of the global river volume[10]. About 60% of this wastewater is treated before 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 (for example, toxins, pesticides, hormones and 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 our understanding of the microbial ecology of activated sludge[14–16], the global picture of microbial diversity and distribution remains elusive. This information is essential for 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 for 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 to systematically collect and analyse activated sludge microbiomes. We

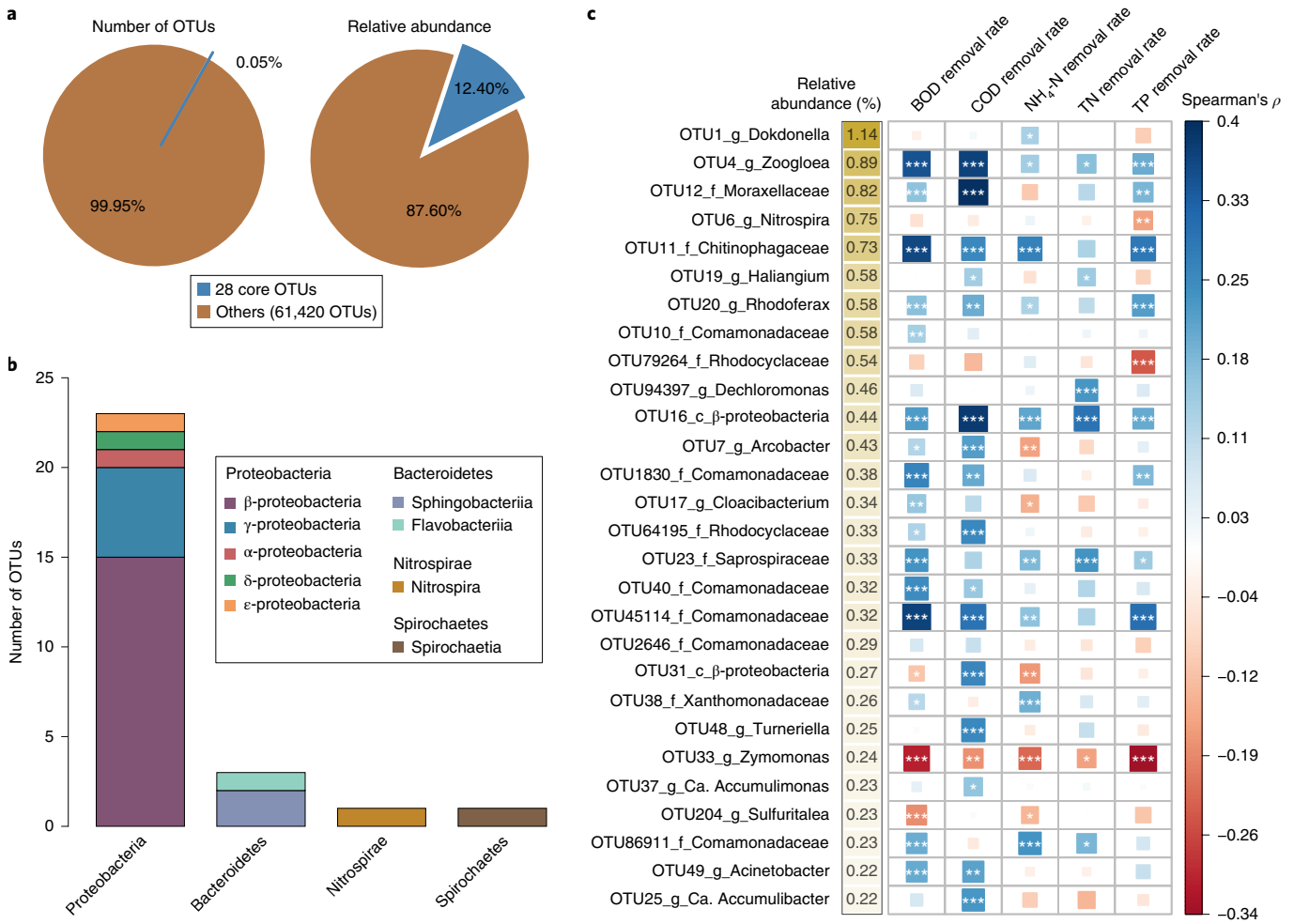

Fig. 2 | Abundance, composition and functional importance of the global core OTUs in activated sludge. a, Percentage and relative abundance of the global core OTUs versus the remaining microbial OTUs. In total, 0.05% (28 out of 61,448 OTUs) were identified as abundant and ubiquitous across WWTPs at the global scale, which accounted for on average 12.4% of the 16S rRNA gene sequences in an activated sludge sample. **b**, The taxonomic composition of the global core OTUs at the phylum and class level. **c**, Activated sludge functions were calculated as the removal rate of organic carbon (BOD removal, 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 colour 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 sizes of the coloured boxes indicate correlation strengths. The asterisks denote the significance levels (two-sided) of the Spearman's rank correlation coefficients ($n = 1{,}186$ biologically independent samples). ***$P < 0.001$, ** $P < 0.01$ and * $P < 0.05$. 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 controlling procedure ($n = 14{,}235$ pairwise cases). Only global core OTUs are shown, with their mean relative abundance indicated on the left of the heatmap.

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 the following: (1) What is the extent of global diversity of activated sludge microbial communities? (2) Does a core microbiome exist in activated sludge processes across different continents? (3) Do activated sludge microbiomes show a latitudinal diversity gradient (LDG)? (4) Is microbial biodiversity important for function in activated sludge processes? (5) 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 the 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

A 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 on the basis of two parameters[19,20]. One is the total number of individuals (N_T), which was estimated as $4–6 \times 10^{23}$ bacteria in the global activated sludge community, based on published data[9]. The other is the quantity of the most abundant taxa (N_{max}), which can be estimated

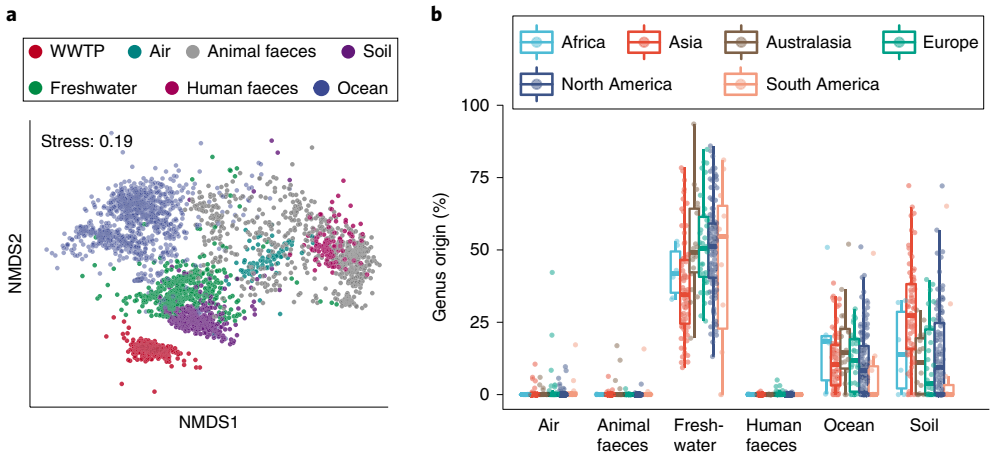

Fig. 3 | Comparing bacterial community compositions across continents and with other habitats. a, NMDS analysis showing that activated sludge of WWTPs harbour a unique microbiome compared with other habitats. For comparison, we merged our OTU table ($n=269$ WWTPs) with that released by the EMP[5], which contained thousands of bacterial communities from various habitats such as soil ($n=338$ samples), ocean ($n=969$ samples), fresh water ($n=447$ samples), air ($n=81$ samples), human faeces ($n=99$ samples) and animal faeces ($n=622$ 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 faeces, fresh water, ocean and soil, as determined by SourceTracker. In the boxplots, hinges show the 25th, 50th and 75th percentiles. The upper whisker extends to the largest value no further than 1.5× the interquartile range (IQR) from the upper hinge, where the IQR is between the 25% and 75% quartiles. The lower whisker extends to the smallest value at most 1.5× the IQR from the lower hinge. Sample sizes for WWTPs are $n=6$, 73, 18, 34, 127 and 11 for Africa, Asia, Australasia, Europe, North America and South America, respectively.

based on either our sequence data or the dominance-scaling law[19]. The lognormal model predicts $1.1\pm0.07\times10^9$ species in activated sludge systems globally, with the N_{max} at 1.2% of the N_T based on our sequence data. The number of species increases only slightly, to $2.0\pm0.2\times10^9$ species, using $N_{max}=0.4\times N_T^{0.93}$ from the dominance-scaling law[19] (Fig. 1c). The estimates of bacterial richness of global activated sludge are only about one order of magnitude lower than that of the global ocean microbiome[19] (~10^{10}), even though the oceans of the world 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 exist in Danish[14] and Asian[15] WWTPs, but less than 10% of the genera overlap. Thus, a global core cannot be established from these regional studies.

At the global scale, occupancy-frequency and occupancy-abundance analyses revealed a hyper-dominant pattern (Supplementary Fig. 1a) in which the 866 most abundant operational taxonomic units (OTUs; 1.39% of the total OTU number) accounted for 50.06% of the total abundance. Similar hyper-dominance patterns were observed in other macrobiological[21] and microbiological communities[22].

A core bacterial community was determined based on the abundance and occurrence frequency of OTUs (see Methods for details). About 0.05% (28 OTUs) constituted a global core that accounted for $12.4\pm0.2\%$ (mean ± s.e.m.) of the sequences in activated sludge samples (Fig. 2a; Supplementary Table 3). Most (82%) of the core community members belonged to *Proteobacteria*, with 15 OTUs classified as β-*Proteobacteria* (Fig. 2b). The most abundant OTU, accounting for $1.14\pm0.05\%$ of the sequence abundance in activated sludge samples and occurring in 85% of all samples, was 99% similar to the γ-proteobacterium *Dokdonella kunshanensis DC-3*[23]. The second most abundant OTU ($0.89\pm0.06\%$ in relative abundance and occurring in 96% of all samples) belonged to *Zoogloea*, a dominant genus in activated sludge communities[15], with *Zoogloea ramigera* known to enhance the flocculation of activated sludge[24]. A *Nitrospira* OTU (OTU_6) was also identified as a core taxon, reflecting its importance for nitrite oxidation or complete ammonia oxidation in activated sludge[25,26]. OTU_7 is closely related to *Arcobacter* species, which are highly abundant in raw sewage[27] and include potential pathogens such as *Arcobacter cryaerophilus*, *Arcobacter butzleri* and *Arcobacter skirrowii*[28]. Furthermore, two putative polyphosphate-accumulating organisms (PAOs), a '*Candidatus* Accumulimonas' OTU (OTU_37) and a '*Candidatus* Accumulibacter' OTU (OTU_25), were identified as core taxa, although only 149 out of the 269 sampled WWTPs operate as enhanced biological P removal (EBPR) systems. Apparently, *C.* Accumulimonas and *C.* 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. A previous study[14] has shown that *Nitrotoga* rather than *Nitrospira* are primary nitrite oxidizers in Danish WWTPs. Another study[30] found low abundances of both *Nitrotoga* and *Nitrospira* in a pilot-scale EBPR treatment plant, but *Nitrotoga* maintained high potential activities based on high small subunit rRNA/rDNA ratios. Regarding PAOs, we identified *C.* Accumulimonas and *C.* 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 faeces, 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 faeces and the ocean, possibly due to the 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

LDGs, whereby species richness tends to decrease as latitude increases[32], are well documented in plant and animal ecology[33].

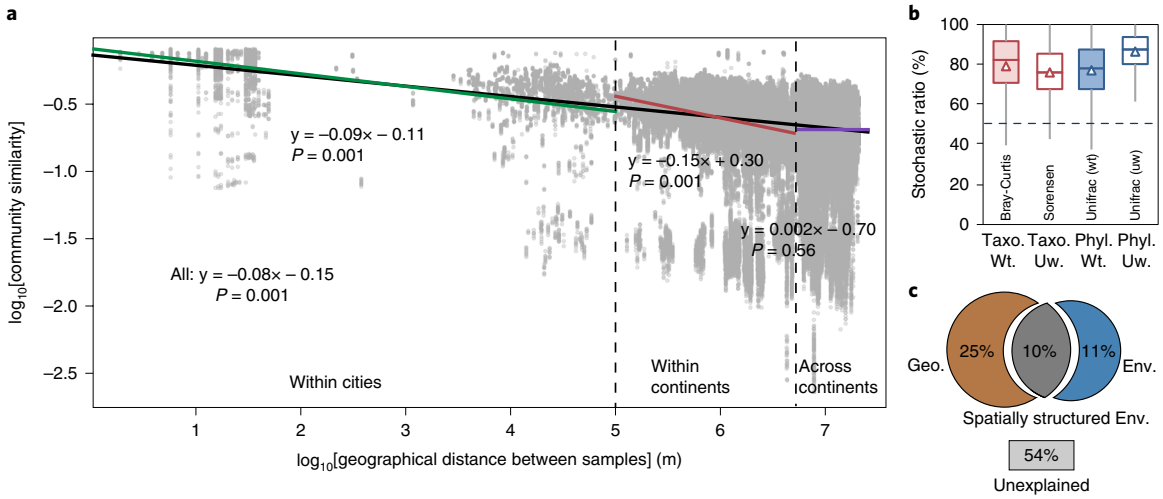

Fig. 4 | Spatial turnover of the activated sludge bacterial communities. a, DDRs based on Bray–Curtis similarity. The black line denotes the least-squares linear regression across all spatial scales ($n = 702{,}705$ pairwise distances). Coloured lines denote separate regressions within cities ($n = 9{,}753$ 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 using matrix permutation tests. **b**, The 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 geographical distance (Geo.) and environmental variables (Env.) to the community variations based on Bray–Curtis distance.

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 LDGs.

We examined the relationship between OTU richness and latitude. OTU richness peaked at intermediate latitude, with a mean air temperature of ~15 °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 ($R^2 < 0.001$; Supplementary Fig. 2b) and mixed liquid temperature ($R^2 = 0.03$; Supplementary Fig. 2c).

Community structure across continents

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), permutational multivariate analysis of variance (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 structures, 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). The activated sludge bacterial communities exhibited higher similarity to those of freshwater and soil than to other environments (Fig. 3a); however, they harboured a unique microbiome that was 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 fresh water, 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 fresh water than the others. Activated sludge and fresh water 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 the geographical distance increases. Consistent with results in other domains[37], we hypothesized that (1) the slope of the DDR curve would vary over local, regional and global scales, and (2) 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 to −0.16 for taxonomic similarity indices, and −0.03 to −0.09 for phylogenetic similarity indices) were significantly ($P = 0.001$) steeper (more than two times) than the overall slopes for all similarity metrics (Supplementary Table 9). Countering our second hypothesis, the overall spatial turnover rates of the activated

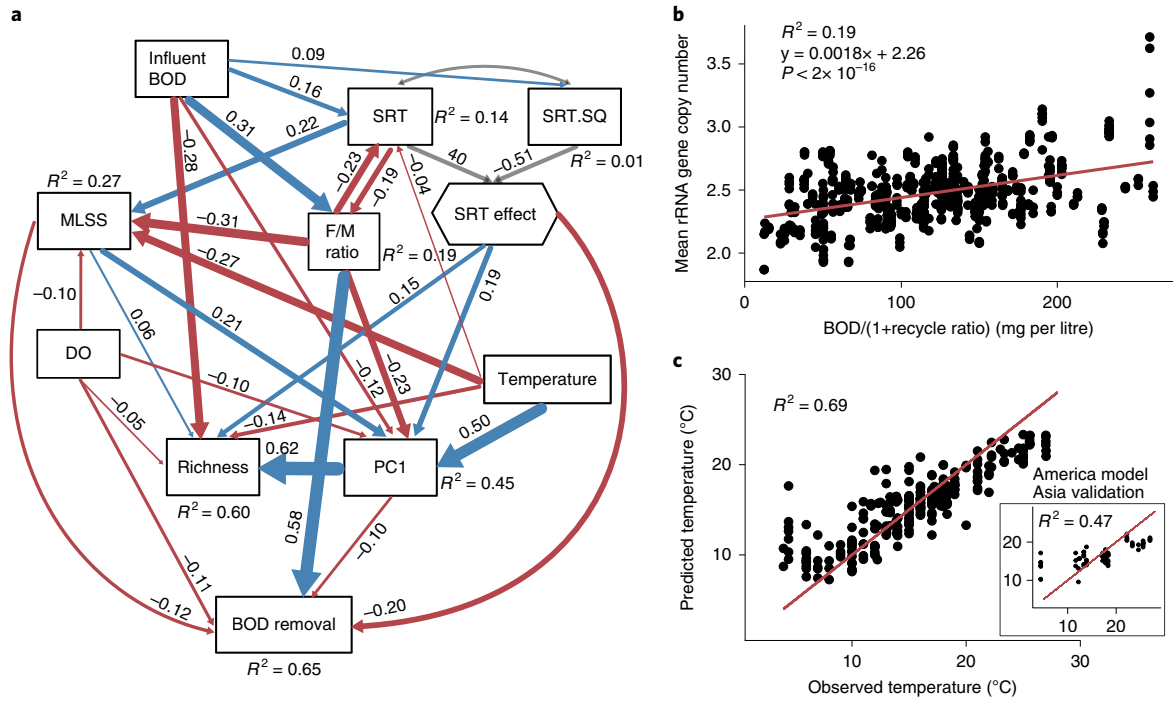

Fig. 5 | Environmental drivers of the activated sludge community composition. a, SEM showing the relationships among environmental variables, community composition and WWTP functioning. The composite variable of 'SRT effect' was constructed as a linear combination of SRT and the square of SRT (SRT.SQ). The community composition is represented by the PC1 from the Bray–Curtis distance-based principal coordinate analysis. DO, dissolved oxygen. 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). The arrow width is proportional to the strength of the relationship. R^2 represents the proportion of variance explained for every dependent variable. Model $\chi^2 = 13.92$, d.f. = 12, $P = 0.31$, $n = 1{,}186$ biologically independent samples; root mean square error of approximation = 0.012 with probability of a close fit = 1.00. **b**, The average rRNA gene copy number of the community increased with the influent BOD/(1 + recycle ratio), which approximates the influent BOD level of the aerobic tank ($n = 641$ biologically independent samples). The red line indicates the regression line. 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 using a random forest analysis ($n = 269$ WWTPs). The red diagonal lines show the theoretical curve for perfect predictions. The inset shows a model trained on data from samples from North and South America to predict the temperature in Asian samples ($n = 73$ WWTPs).

sludge communities were similar to those found in non-flowing natural habitats such as soils[6] and sediments[37].

Linking community structure to function

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 out of the five functions examined. Most of the correlations (81%) were positive (Fig. 2c). Moreover, 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 indicate that the structure of the activated sludge bacterial communities, particularly the dominant populations, is critical for maintaining activated sludge functions.

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

Consistent with our expectation, the activated sludge community composition was significantly correlated with the total P removal rate for the samples from EBPR plants, but not for non-EBPR plants (Supplementary Table 10), as P removal processes in non-EBPR plants are predominantly chemical. The diversity of the three potential PAOs[31] were significantly different ($P < 0.0001$, two-tailed paired t-test between any two organisms), with 8.2 ± 0.2 *C.* Accumulimonas OTUs, 6.6 ± 0.2 *C.* Accumulibacter OTUs

and 3.2 ± 0.1 *Tetrasphaera* OTUs within a typical activated sludge bioreactor. While the relative abundance of *C.* Accumulimonas (0.42 ± 0.06%) was not different from that of *C.* Accumulibacter (0.42 ± 0.04%; two-tailed paired *t*-test, $P = 0.92$), both were more abundant than *Tetrasphaera* (mean relative abundance 0.17 ± 0.02%; two-tailed paired *t*-test, $P < 0.0001$) (Supplementary Table 12).

Stochastic community assembly

Since WWTPs are well-controlled engineered ecosystems, we hypothesized that the activated sludge community assembly has a deterministic nature, and we calculated the null model-based stochastic ratios[41] with taxonomic and phylogenetic metrics. The average stochastic ratios based on these four metrics were all higher than 0.75 (Fig. 4b). This result suggests that stochastic factors are more important than deterministic factors in influencing community composition, at least partially contradicting our hypothesis.

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

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 geographical distance-corrected dissimilarities of community composition with those of environmental variables using the partial Mantel test (Supplementary Fig. 5a; Supplementary Table 13). Overall, the microbial community composition was strongly correlated with absolute latitude, mean annual temperature, solids retention time (SRT, the average time during which activated sludge solids are in the system) and influent COD and BOD concentrations, representing organic matter ($r_m = 0.23$–0.30, $P = 0.001$).

A more in-depth analysis using structural equation modelling (SEM) revealed direct and indirect effects of the environmental drivers (Fig. 5a). Consistent with the Mantel test results, temperature had the strongest direct effects on the first principal component score (PC1) representing the community structure (standardized path coefficient, $\beta = 0.50$, $P < 0.001$). It also had weak negative impacts on species richness ($\beta = -0.14$, $P < 0.001$). This is consistent with previous observations at local[45,46] and 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 ammonium-oxidizing bacteria and nitrite-oxidizing bacteria.

Various biotic and abiotic factors (for example, 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 probably has an impact on bacterial composition through its effect on the F/M ratio ($\beta = 0.31$, $P < 0.001$), which is inversely related to the SRT. Influent BOD is the most influential environmental variable directly related to bacterial richness ($\beta = -0.28$, $P < 0.001$), and the abundance-weighted mean rRNA gene copy number significantly increased with the influent BOD ($R^2 = 0.19$, $P < 0.0001$) (Fig. 5b). All of these results are consistent with the resource-competition theory[48], which predicts that high species diversity occurs with low to intermediate supply of resources, but fast-growing r-strategists outcompete efficient-scavenging K-strategists at high resource levels[49].

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

Conclusions and future perspectives

Through well-coordinated international efforts, we systematically examined the global diversity and biogeography of activated sludge bacterial communities within the context of theoretical ecology frameworks. Our findings enhance our 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 probably by stochastic processes, such as dispersal and drift, which apparently contradicts conventional wisdom. However, deterministic factors (for example, 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. Alternatively, one could alternate organic C loadings and/or operational conditions to manipulate the structure of the activated sludge community to select for the microorganisms that have 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. For example, the relationships among biodiversity, food web interactions, succession, stability and ecosystem functioning.

Methods

Global sampling and metadata collection. The GWMC (http://gwmc.ou.edu/) was initiated in May 2014 as a platform to facilitate international collaboration and communication on research and education for global water microbiome studies. 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 WWTPs by focusing on the activated sludge process. Unlike the 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 a system-level mechanistic understanding of the global diversity and distribution of municipal WWTP microbiomes. WWTPs were selected based on the following criteria: (1) Continental-level geographical locations. Samples were obtained from all continents except for Antarctica, but with a special focus on North America, Asia and Europe (Fig. 1a). Because of the low accessibility, WWTPs in Africa and South America were under-represented. (2) Latitude. To address questions related to the LDG, WWTPs were intensively sampled in North America along the east and west coasts, and Highway 35 and Highway 40 (from east to west) (Fig. 1a), in Asia, Europe and Australia. The WWTPs sampled spanned latitudes from 43.6 °S to 64.8 °N. (3) Climate zones. Since climate could have substantial impacts on microbial communities, the samples covered 17 different climate types (Supplementary Fig. 6). To distinguish independent effects of continents versus climate zones, we increased sampling efforts for climate zones that were present in multiple continents, such as humid subtropical climate. (4) Scales. The samples were collected from very broad spatial scales; that is, global (across 6 continents), regional (for example, individual continents or climate zones) and local (for example, individual cities). Within some cities, multiple WWTPs and multiple samples per WWTP were collected. (5) Wastewater treatment process types. To address the relationship between structure and function for activated sludge, we sampled the aerobic zone of conventional plug flow, oxidation ditch, sequential batch reactors, anaerobic/anoxic/oxic and other activated sludge process types.

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

Sampling was carried out between June and November 2014 in the Northern Hemisphere and between December 2014 and April 2015 in the Southern Hemisphere. The sampling time was generally between 10:00 to 14:00, when the WWTPs were relatively stable under normal conditions. Although we tried to collect the global samples in the same season, seasonal temporal turnover in activated sludge communities could have had some effect on the community variations we observed. Based on limited published work[54,55], such temporal variations should be much smaller than the spatial variations at the global or continental scales. For example, a previous study of the 5-year temporal dynamics of an activated sludge community showed no significant seasonal succession[54]. It also revealed that the activated sludge communities were relatively stable across 3 months, with an average Bray–Curtis distance of 0.45 ± 0.10 (mean ± s.d.) between samples[55]; this variation was smaller than our observed mean variations even at the local city level (0.54 ± 0.19) (Fig. 4a).

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

At each sampling position, approximately 1 litre of mixed liquor was sampled and well mixed, and 40 ml was transferred into a sterile tube. The mixed liquor samples were kept on ice (≤4 °C), transported to a laboratory within 24 h, divided into aliquots and then centrifuged at 4 °C, 15,000 × *g* for 10 min to collect pellets. Sludge pellets were transported (if necessary) with dry ice to the designated laboratories within 48 h and preserved at −80 °C before DNA extraction.

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

Climate type was determined using the Köppen–Geiger climate classification[58]. Gross domestic product and population data were derived from the Brookings analysis of Global Metro Monitor[59]. Variables related to plant design and operation include plant age, design capacity, actual flow rate, volume of aeration tanks, hydraulic retention time (HRT) and SRT. The activated sludge process type, aerator type and coupling with N removal processes (nitrification and denitrification) in the WWTP were also provided by the plant managers where possible. Plant monitoring variables included influent and effluent BOD and COD (representing the organic C level), total nitrogen and total phosphorus (representing nutrient level), ammonium N, and the F/M ratio (indicating the average organic C loading to microorganisms). For sample properties, most plant managers provided the yearly average value of mixed liquor suspended solids (MLSS), indicating the concentration of biomass in the activated sludge, dissolved oxygen, pH and mixed liquid temperature; some provided the measured values when sampling.

Activated sludge performance was calculated as the specific removal rates (g per g biomass per day) of organic C (BOD and COD), nutrients (total nitrogen and total phosphorus) and ammonium nitrogen (NH_4-N) as follows:

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

The WWTPs represent diverse geographies and a large range of climatic conditions, operation parameters and chemical conditions across and within continents (Supplementary Fig. 7). For instance, the average influent BOD ranged from 30 to 1,000 mg per litre. Such a broad range of diverse parameters is critical for disentangling the 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 X.W.'s Laboratory at Tsinghua University for DNA extraction. All other samples, including samples from Europe collected by T.C. at Newcastle University, were shipped to J.Z.'s Laboratory at the University of Oklahoma for DNA extraction. Owing 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 kits. For each sample, a pellet from 3 ml of mixed liquor was used. In addition to the manufacturer's protocol, we always placed 12 bead tubes evenly on the vortexer and vortexed at maximum speed for 10 min to minimize the lysis efficiency difference between samples. All DNA samples were processed at The University of Oklahoma for sequencing.

DNA quality for all samples was evaluated using a NanoDrop spectrophotometer (NanoDrop Technologies) at the University of Oklahoma. Final DNA concentrations were quantified using PicoGreen and a FLUO star Optima instrument (BMG Labtech). 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 EMP[61]. In silico primer coverage analysis using SILVA TestPrime v.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.

To mitigate quantitative problems associated with amplicon sequencing[52], the 16S rRNA gene fragments were amplified from community DNA samples (10 ng) with two-step PCR using lower numbers of amplification cycles (10 and 20 cycles for the first and second step, respectively). The two-step PAS approach provides the following advantages: lower amplification biases, better sequence-read quality, higher effective sequence read numbers and length, and lower sequencing errors[60]. All samples were sequenced using the same MiSeq instrument at the Institute for Environmental Genomics, University of Oklahoma. Generally, about 400 samples were combined for each round of MiSeq sequencing. Since the numbers of sequence reads varied substantially from sample to sample, most samples were sequenced more than once (for example, 19% twice, 33% three times, 43% more than times) to meet the target number of about 30,000 sequencing reads per sample, as determined in our previous analysis[63].

The numbers of sequences (reads) per sample ranged from 25,631 to 351,844 (Supplementary Table 5), and a total of 96,148 OTUs were obtained. About 1.3% of these OTUs were from archaea, which accounted for 0.13% of the total abundance.

The choice of the PCR primer pair 506F/806R (which was also used in the EMP) is highly likely to have strongly influenced this low archaeal abundance due to the much lower coverage of the primers of archaeal 16S rRNA genes compared to the bacterial counterparts. Because of the low archaeal abundance, the term 'bacteria' is used for simplicity. Also, the terms microbiome and microbial (or bacterial) community are used interchangeably.

Raw sequence data were processed as previously described[35], except for OTU generation by UPARSE[64] at the 97% similarity threshold, resulting in 96,148 OTUs. We define OTUs (based on 97% sequence similarity) for bacterial and archaeal phylotypes. Although there is potential misconnection between OTUs and microbial species[65], we use this popular definition for simplicity, and it allows comparison with previous studies of other systems. The representative sequences were aligned using Clustal Omega v.1.2.2[66] for constructing the phylogenetic tree using FastTree2 v.2.1.10[67]. OTUs were taxonomically annotated using RDP Classifier[68] with a confidence cut-off of 80%, using the MiDAS database (v.2.1), which specifically provides a curated taxonomy for abundant and functionally important microorganisms in activated sludge[69]. After removal of the global singletons[64], the sequence number in each sample was rarefied to the same depth (25,600 sequences per sample), resulting in 61,448 OTUs overall, which were used in subsequent comparative analyses.

Although our sequencing depths were considerably higher than those in many similar studies[70], rarefaction curves (Supplementary Fig. 2d,e) of activated sludge microbial communities indicated that additional rare taxa were probably present in individual samples. Nevertheless, pooling all sequences gave a sufficient number for estimating global- and continent-level diversity of activated sludge microbial communities (Supplementary Fig. 2f,g). The global OTU richness per sample was $2{,}309 \pm 559$ (Supplementary Table 5). Besides richness, we also calculated other α-diversity indices on a global and regional scale (Supplementary Table 5).

The rRNA operon copy number for each OTU was estimated through the rrnDB database based on its closest relatives with a known rRNA operon copy number[71]. The abundance-weighted mean rRNA operon copy number was then calculated for each sample as described previously[49].

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

SAD fitting. We compared the SAD of each sample, based on the rank-abundance distribution, with predictions from Poisson lognormal, log-series, Broken-stick and Zipf models. Although numerous SAD models are available, lognormal and log-series have been the most successful in predicting SADs, and they are the standards for testing other models[18]. While the logseries model is well supported by macroecological studies, the Poisson lognormal model is more 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 a previously described code[18].

Estimation of global bacterial diversity of WWTPs. We used the methods described in previous studies[19,20] 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 as follows:

$$S_T = \frac{\sqrt{\pi}}{a}\exp\left\{\left(a\log_2\left(\sqrt{\frac{N_{max}}{N_{min}}}\right)\right)^2\right\} \quad (1)$$

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

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

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 a previous study[76], suggest that about 330 $km^3\,year^{-1}$ of municipal wastewater are produced globally, of which 60% is treated[9]. Assuming that they are all treated in WWTPs, then about 0.54 km^3 of municipal wastewater is treated by WWTPs globally per day. The total effective volume (V) of aerobic tanks of WWTPs can be estimated as follows:

$$V = Q \times \mathrm{HRT} \quad (3)$$

where Q is the influent flow rate ($m^3\,day^{-1}$) and the HRT of the aerobic tank is measured in days. Our dataset indicates that the average HRT of aerobic tanks is 9.8 h ($\pm$0.3 s.e.). Thus, the total effective volume is estimated as 0.22 ± 0.007 km^3. The total number of cells in activated sludge is about $2.3 \pm 0.4 \times 10^9$ ml^{-1} (ref. [77]); thus, the N_T (global activated sludge bacterial abundance) is about $4.0–6.1 \times 10^{23}$.

We then estimated the N_{max} based on the ratio of N_{max} to N_T of our sequencing data, that is, the relative abundance of the most abundant OTU, or using scaling law[19]. The knowledge of N_T, N_{max} and N_{min} allows equation (2) to be solved numerically for the parameter a and, subsequently, for S_T using equation (1).

Using the same method, we estimated the total bacterial richness of individual WWTPs, along with WWTPs in the United States and China. The volume of aerobic tanks of a WWTP in Beijing, China is 10,000 m^3, making the total number of cells about $2.3 \pm 0.4 \times 10^{19}$. N_T of WWTPs in the United States and China were estimated based on their published data of the amount treated[78,79], with activated sludge harbouring similar numbers of species for the United States (4.6×10^8 to 1.1×10^9) and China (3.9×10^8 to 1.0×10^9). N_{max} was further estimated based on our 16S rRNA gene sequencing data or using a scaling law[19]. The total bacterial richness estimates of individual human gut, individual cow rumen, global ocean and Earth were taken from a previous study[19].

Core community determination. A global-scale core microbial community was determined based on multiple reported measures. First, 'overall abundant OTUs' were filtered out according to mean relative abundance (MRA) across all samples[80]. Previous studies used different criteria (for example, MRA > 1%[30,81] or 0.1%[82,83]) without any objective or standard rule. Thus, we selected all top 0.1% OTUs (62) as overall abundant OTUs. Their MRA was higher than 0.2%, within the range of reported criteria. Second, 'ubiquitous OTUs' were defined as OTUs with an occurrence frequency in more than 80% of all samples[84]. Finally, 'frequently abundant OTUs' were selected based on their relative abundances with a sample. In each sample, the OTUs were defined as abundant when they had a higher relative abundance than other OTUs and made up the top 80% of the reads in the sample[14]. A frequently abundant OTU was defined as abundant in at least half of the samples, which is stricter than the reported criterion (10 out of 26 samples)[14]. Since the above three measures are complementary to one another when defining core community, only OTUs fulfilling all three criteria were defined as the global scale core bacterial community.

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

Comparison of bacterial community composition of WWTPs to natural habitats and source tracking. We downloaded the OTU table of 16S rRNA gene amplicon studies from the EMP (ftp://ftp.microbio.me/emp/release1/otu_tables/closed_ref_greengenes/emp_cr_gg_13_8.subset_5k.biom)[5]. This table was generated using closed reference against Greengenes 13.8 and contained 5,000 global samples from multiple habitats. To compare community compositions at the OTU level, our activated sludge OTUs were repicked using closed reference against Greengenes 13.8, which picked 68.1% of the sequences. This OTU table was then merged with the EMP OTU table. To give relatively equal representation of samples across environments, we further collapsed our activated sludge samples at the plant level by summing the abundance of each OTU across samples of the same plant, resulting in 269 activated sludge samples. Our activated sludge samples and the EMP samples from fresh water (including that from fresh water and freshwater biofilm), ocean (including that from sea water and biofilm), animal faeces, human faeces, soil and air were selected from the merged OTU table. We then subsampled to 10,000 sequences per sample. To compare microbial community compositions across habitats, the nonmetric multidimensional scaling (NMDS) analysis was performed using the Bray–Curtis dissimilarity matrix.

The proportion of each activated sludge microbiota attributable to fresh water, soil, ocean, animal and human faeces, and air at the genus level were estimated using SourceTracker[39], which was run through QIIME with default settings and using activated sludge microbiota as the sink and those in other habitats as sources. Genera detected in less than 1% of the samples were filtered out before source-tracking modelling.

Diversity analyses using 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 the Shannon index, the 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 using 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 the main climate group level, which includes the following five groups: A (tropical), B (arid), C (temperate), D (cold) and E (polar)[58]. The activated sludge process types were classified into the following nine general groups: complete mix, conventional plug flow, sequential batch reactors, anaerobic/anoxic/oxic, anoxic/oxic, oxidation ditch, contact stabilization, pure oxygen and extended aeration.

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

Estimating stochasticity of community assembly. We assessed community assembly stochasticity using a null-model-based index. The stochasticity ratio has been described previously[41,89]. Since null-model algorithms usually require a high number of replicates, we selected 71 cities, each of which had more than 9 samples; we randomly drew 9 samples from each city to make sampling even. We calculated the stochasticity ratio using taxonomic and phylogenetic metrics. Whether using the Bray–Curtis (abundance-weighted) or Sorensen (unweighted) model, the stochasticity ratio was calculated based on typical null-model algorithms for taxonomic metrics[90,91]. When using weighted and unweighted Unifrac, the stochasticity ratio was calculated based on typical null-model algorithms for phylogenetic metrics[91,92]. Samples within each city were considered sharing the same regional species pool in null-model algorithms.

Partitioning the environment and distance effect. To provide a quantification of the relative contribution of the environment effect versus the distance effect on β-diversity, we performed a VPA based on 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's $r = -0.25$), and it indicated a low level of collinearity among these variables. MRM was performed in different spatial scales. Geographical 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 nonsignificant variables from this initial MRM test, and then re-ran 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 following four components of variations as described elsewhere[94]: (1) pure environmental variation $= R^2_T - R^2_G$; (2) pure geographical distance $= R^2_T - R^2_E$; (3) spatially structured environmental variation $= R^2_G + R^2_E - R^2_T$; and (4) unexplained variation $= 1 - R^2_T$.

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 PC1 of principal coordinate analysis based on Bray–Curtis distance. We first considered a full model that included all reasonable pathways, and then we sequentially eliminated nonsignificant 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. 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 variables and compositional data using the randomForest R package[96]. We used OTUs as predictors and environmental variables 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 that 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 dataset for training, and we subsequently trained new random forests for each plant using customized training sets that excluded plants within a defined radius of the target plant. The size of this radius ranged from 0 to 5,000 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).

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

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 website (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.

Received: 7 September 2018; Accepted: 8 March 2019;
Published online: 13 May 2019

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**, e02288-14 (2015).
5. Thompson, L. R. et al. A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (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, Engineering, and 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* (eds Drechsel, P., Qadir, M. & Wichelns, D.) 15–38 (Springer, 2015).
10. Gleick, P. H. in *Encyclopedia of Climate and Weather* (ed. Schneider, S. H.) 817–823 (Oxford Univ. Press, 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 (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 of *Nitrospira*-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**, 525 (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 (2008).
35. Zhou, J. et al. Temperature mediates continental-scale diversity of microbes in forest soils. *Nat. Commun.* **7**, 1208 (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 (2011).
38. Zhou, J. et al. Stochastic assembly leads to alternative communities with distinct functions in a bioreactor microbial community. *mBio* **4**, e00584-12 (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–e00017 (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**, 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 Univ. 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 (2017).
50. Pedrós-Alió, C. & Manrubia, S. The vast unknown microbial biosphere. *Proc. Natl Acad. Sci. USA* **113**, 6585–6587 (2016).
51. Zhou, J. et al. Random sampling process leads to overestimation of β-diversity of microbial communities. *mBio* **4**, e00324 (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. Sc.* **4**, 439–473 (2007).
59. Berube, A., Leal Trujillo, J., Ran, T. & Parilla, J. *Global Metro Monitor* (Brookings, 2015); https://www.brookings.edu/research/global-metro-monitor/
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 (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**, e02149-17 (2018).
66. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence alignments using Clustal Omega. *Mol. Syst. Biol.* **7**, 539 (2011).
67. Price, M. N., Dehal, P. S. & Arkin, A. P. FastTree 2—approximately maximum-likelihood trees for large alignments. *PLoS ONE* **5**, e9490 (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**, bax016 (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.* **46**, D593–D598 (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 Gobal Information System on Water and Agriculture. Wastewater Section* (FAO, 2014); http://www.fao.org/nr/water/aquastat/wastewater/index.stm
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. *The Sources and Solutions: Wastewater* (United States Environmental Protection Agency, 2018).
79. Chan, W. *Wastewater: Good To The Last Drop* (China Water Risk, 2017); http://chinawaterrisk.org/resources/analysis-reviews/wastewater-good-to-the-last-drop/
80. Hanski, I. Dynamics of regional distribution: the core and satellite species hypothesis. *Oikos* **38**, 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 (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.* **9**, 683–695 (2015).
85. Kembel, S. W. et al. Picante: R tools for integrating phylogenies and ecology. *Bioinformatics* **26**, 1463–1464 (2010).
86. Oksanen, J. et al. vegan: Community Ecology Package. R 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 version 1 (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). *J. Stat. Soft.* **48**, 1–36 (2012).
96. Liaw, A. & Wiener, M. Classification and regression by randomForest. *R News* **2**, 18–22 (2002).

Acknowledgements

The authors thank T. Allen, A. Al-Omari, R. Bart, D. Crowley, G. Harwood, T. Hensley, S.-J. Huitric, M. M. L. Martins, A. Mena, B. Pathak, S. Pereira, D. E. Sauble, M. Taylor, P. Truong, D. VanderSchuur, A. Vieira and D. Zambrano for helping with sampling and metadata collection. This work was supported by the 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. Lin.W. and B.Z. were generously supported by the China Scholarship Council (CSC). J.Z. (jzhou@ou.edu) and D.N. (ningdaliang@ou.edu) serve as GWMC contacts.

Author contributions

All authors contributed experimental assistance and intellectual input to this study. The original concept was conceived by J.Z. Experimental strategies and sampling design were developed by J.Z., X.W., T.P.C., Q.H., Z. He. and D.N. Sample collections were coordinated by Q.H., D.N., X.W., T.P.C., B.Z., M.B., G.F.W., J.Z. and other GWMC members. J.D.V.N and D.N. managed shipping. Y. Li., B.Z., ZX.L., D.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-H., A.C., P. Bovio. and D.A.) did DNA extraction. P.Z. performed DNA sequencing with the help from Liy.W. Data analyses were performed by Lin.W., D.N., J.Z., B.Z., X.S., Q.Z., F.L., N.X. and R.T. with help from Y.D., Q.T., T.Z., Ya.Z and A.W. The manuscript was written by Lin.W., J.Z. and D.N. 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., X.W., Z. He. and Y.Y.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41564-019-0426-5.

Reprints and permissions information is available at www.nature.com/reprints.

Correspondence and requests for materials regarding analysis, synthesis, and reprints should be addressed to J.Z.; and regarding experimental design and sampling should be addressed to X.W., T.P.C. or Q.H.

Publisher's note: Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Global Water Microbiome Consortium

Dany Acevedo[30], Miriam Agullo-Barcelo[18], Pedro J. J. Alvarez[20], Lisa Alvarez-Cohen[26,27], Gary L. Andersen[31,32], Juliana Calabria de Araujo[33], Kevin Boehnke[34], Philip Bond[18], Charles B. Bott[35], Patricia Bovio[36], Rebecca K. Brewster[34], Faizal Bux[37], Angela Cabezas[36], Léa Cabrol[38,39], Si Chen[23], Craig S. Criddle[21], Ye Deng[12], Claudia Etchebehere[36], Amanda Ford[35], Dominic Frigon[40], Janeth Sanabria Gómez[30], James S. Griffin[41], April Z. Gu[42], Moshe Habagil[43], Lauren Hale[2], Steven D. Hardeman[44], Marc Harmon[45], Harald Horn[46], Zhiqiang Hu[47], Shameem Jauffur[40,48], David R. Johnson[49], Jurg Keller[18], Alexander Keucken[43,50], Sheena Kumari[37], Cintia Dutra Leal[33], Laura A. Lebrun[51], Jangho Lee[52], Minjoo Lee[52], Zarraz M. P. Lee[53], Yong Li[4], Zhenxin Li[7], Mengyan Li[20], Xu Li[54], Fangqiong Ling[55], Yu Liu[53,56], Richard G. Luthy[21], Leda C. Mendonça-Hagler[57], Francisca Gleire Rodriguez de Menezes[58], Arthur J. Meyers[59], Amin Mohebbi[54,60], Per H. Nielsen[19], Daliang Ning[2,3,1], Adrian Oehmen[61], Andrew Palmer[62], Prathap Parameswaran[28], Joonhong Park[52], Deborah Patsch[49], Valeria Reginatto[63], Francis L. de los Reyes III[64], Bruce E. Rittmann[28], Adalberto Noyola Robles[65], Simona Rossetti[66], Xiaoyu Shan[1], Jatinder Sidhu[62], William T. Sloan[55], Kylie Smith[62], Oscarina Viana de Sousa[58], David A. Stahl[25], Kyle Stephens[67], Renmao Tian[2], James M. Tiedje[22], Nicholas B. Tooker[68], Qichao Tu[12], Joy D. Van Nostrand[2], Daniel De los Cobos Vasconcelos[65], Julia Vierheilig[9,10], Michael Wagner[9], Steve Wakelin[69], Aijie Wang[13], Bei Wang[70], Joseph E. Weaver[64], George F. Wells[11], Stephanie West[46], Paul Wilmes[51],

Sung-Geun Woo[21], Linwei Wu[1,2], Jer-Horng Wu[71], Liyou Wu[2], Chuanwu Xi[34], Naijia Xiao[2,3], Meiying Xu[72], Tao Yan[73], Yunfeng Yang[1], Min Yang[74], Michelle Young[28], Haowei Yue[1], Bing Zhang[1,2], Ping Zhang[2,5], Qiuting Zhang[1], Ya Zhang[2], Tong Zhang[15], Qian Zhang[73], Wen Zhang[75], Yu Zhang[74], Hongde Zhou[70], Jizhong Zhou[1,2,33], Xianghua Wen[1], Thomas P. Curtis[36], Qiang He[23,24], Zhili He[16,17] and Matthew Brown[6]

[30]Environmental Microbiology and Biotechnology Laboratory, Engineering School of Environmental and Natural Resources, Engineering Faculty, Universidad del Valle–Sede Meléndez, Cali, Colombia. [31]Department of Environmental Science, Policy, and Management, University of California, Berkeley, CA, USA. [32]Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA. [33]Universidade Federal de Minas Gerais, Departamento de Engenharia Sanitária e Ambiental, Belo Horizonte, Brazil. [34]Department of Environmental Health Sciences, The University of Michigan, Ann Arbor, MI, USA. [35]Hampton Roads Sanitation District (HRSD), Virginia Beach, VA, USA. [36]Microbial Ecology Laboratory, Microbial Biochemistry and Genomics Department, Biological Research Institute "Clemente Estable", Montevideo, Uruguay. [37]Institute of Water and Wastewater Technology, Durban University of Technology, Durban, South Africa. [38]Aix-Marseille University CNRS IRD, MIO UM110 Mediterranean Institute of Oceanography, Marseille, France. [39]Escuela de Ingenieria Bioquimica, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile. [40]Microbial Community Engineering Laboratory, Department of Civil Engineering and Applied Mechanics, McGill University, Montreal, Québec, Canada. [41]Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA. [42]School of Civil and Environmental Engineering, Cornell University, Ithaca, NY, USA. [43]Vatten and Miljö i Väst AB (VIVAB), Falkenberg, Sweden. [44]Norman Water Reclamation Facility, Norman, OK, USA. [45]Golden Heart Utilities, Fairbanks, AK, USA. [46]Karlsruhe Institute of Technology, Engler-Bunte-Institut, Water Chemistry and Water Technology, Karlsruhe, Germany. [47]Department of Civil and Environmental Engineering, University of Missouri, Columbia, MO, USA. [48]Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Québec, Canada. [49]Department of Environmental Microbiology, Eawag, Dübendorf, Switzerland. [50]Water Resources Engineering, Faculty of Engineering, Lund University, Lund, Sweden. [51]Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg. [52]Department of Civil and Environmental Engineering, Yonsei University, Seoul, South Korea. [53]Advanced Environmental Biotechnology Centre, Nanyang Environment and Water Research Institute, Nanyang Technological University, Singapore, Singapore. [54]Department of Civil Engineering, University of Nebraska, Lincoln, NE, USA. [55]Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow, UK. [56]School of Civil and Environmental Engineering, Nanyang Technological University, Singapore, Singapore. [57]Plant Biotechnology Program, Federal University of Rio de Janeiro, UFRJ, Rio de Janeiro, Brazil. [58]Federal University of Ceará, UFC, Ceará, Brazil. [59]University of Tennessee, Center for Environmental Biotechnology, Knoxville, TN, USA. [60]Department of Civil Engineering, Construction Management and Environmental Engineering, Northern Arizona University, Flagstaff, AZ, USA. [61]UCIBIO, REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal. [62]CSIRO Land and Water, Ecosciences Precinct, Dutton Park, Queensland, Australia. [63]Departamento de Química, Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto—FFCLRP, Ribeirão Preto, Brazil. [64]Department of Civil, Construction, and Environmental Engineering, North Carolina State University, Raleigh, NC, USA. [65]Grupo de Investigación en Procesos Anaerobios, Instituto de Ingeniería, Universidad Nacional Autónoma de México, México, Mexico. [66]CNR-IRSA, National Research Council, Water Research Institute, Rome, Italy. [67]Tryon Creek and Columbia Blvd Wastewater Treatment Plants, Bureau of Environmental Services, City of Portland, OR, USA. [68]Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA. [69]Scion Research, Christchurch, New Zealand. [70]School of Engineering, University of Guelph, Guelph, Ontario, Canada. [71]Department of Environmental Engineering, National Cheng Kung University, Tainan City, China. [72]State Key Laboratory of Applied Microbiology Southern China, Guangdong Provincial Key Laboratory of Microbial Culture Collection and Application, Guangdong Institute of Microbiology, Guangzhou, China. [73]Department of Civil and Environmental Engineering, University of Hawaii at Manoa, Honolulu, HI, USA. [74]State Key Laboratory of Environmental Aquatic Chemistry, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China. [75]Department of Civil Engineering, University of Arkansas, Fayetteville, AR, USA.

nature research

Corresponding author(s): Jizhong Zhou

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see Authors & Referees and the Editorial Policy Checklist.

Statistical parameters

When statistical analyses are reported, confirm that the following items are present in the relevant location (e.g. figure legend, table legend, main text, or Methods section).

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	An indication of whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☐	☒	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☐	☒	A description of all covariates tested
☐	☒	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☐	☒	A full description of the statistics including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☐	☒	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☐	☒	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☐	☒	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated
☐	☒	Clearly defined error bars *State explicitly what error bars represent (e.g. SD, SE, CI)*

Our web collection on statistics for biologists may be useful.

Software and code

Policy information about availability of computer code

Data collection	No software used for data collection
Data analysis	UPARSE, Clustal Omega (v1.2.2), FastTree2 (v2.1.10) and RDP Classifier were used to process the sequencing data; R (v3.4.6) was used for statistical analyses.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors/reviewers upon request. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on 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). R codes on the statistical analyses are available from the corresponding authors upon reasonable request.

Field-specific reporting

Please select the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☐ Life sciences ☐ Behavioural & social sciences ☒ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/authors/policies/ReportingSummary-flat.pdf

Ecological, evolutionary & environmental sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	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/). 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 (AS) process. The main goal of this study was to provide system-level mechanistic understanding of global diversity and distribution of AS microbiomes from WWTPs.
Research sample	The research sample was the activated sludge in the aerobic zone of WWTPs. The 16S rRNA sequence data were obtained from the activated sludge samples.
Sampling strategy	WWTPs were selected considering their representation in continental-level geographic locations, latitude, climate zones and wastewater treatment process types (see Methods for details). In each plant, we collected at least three AS samples, generally from three different positions (the front, middle, and end part) of the aerobic zone in each aeration tank. In a few cases (3.3% plants), where only one sampling position was applicable, three samples were taken in sequence with at least 30-min interval. In total, 1,186 AS samples were collected from 269 WWTPs in 86 cities, 23 countries, and 6 continents.
Data collection	Community DNAs were extracted from the 1186 AS samples using a standardized approach. The V4 region of the 16S rRNA gene was amplified and deeply sequenced in one laboratory using the same method for all samples, to obtain the 16S rRNA sequence data. The metadata were provided by plant managers and/or investigators. Along with the sludge samples, we collected metadata (e.g., chemical properties, operation conditions, process type) from each plant using a standard sampling data sheet, which ensured that the data from all plants was in the same format.
Timing and spatial scale	The sampling was carried out in June to November 2014 in the Northern Hemisphere and December 2014 to April 2015 in the Southern Hemisphere. The sampling time was generally between 10:00 am to 2:00 pm, when the WWTPs were relatively stable under normal conditions. The samples were collected from very broad spatial scales: global (across 6 continents), regional (e.g., individual continents or climate zones), and local (e.g., individual cities). In total, 1,186 AS samples were collected from 269 WWTPs in 86 cities, 23 countries, and 6 continents.
Data exclusions	No data were excluded.
Reproducibility	No laboratory manipulation experiments of biological processes were conducted - rather, and as stated above, our analyses are based on one-time survey conducted on AS of municipal WWTPs across a wide range of geographic locations.
Randomization	No experiments per se were conducted, there was thus no experimental group allocation. There was no further group partitioning of data beyond the natural groupings associated with geography.
Blinding	Blinding was not relevant to our study, because all available data were used (our study did not perform an experiment).

Did the study involve field work? ☐ Yes ☒ No

Reporting for specific materials, systems and methods

Materials & experimental systems

n/a	Involved in the study
☒	☐ Unique biological materials
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology
☒	☐ Animals and other organisms
☒	☐ Human research participants

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging