Global Diversity and Biogeography of the Bacterial Communities That Clean Our Wastewater

Linwei Wu1,2,*, Daliang Ning2,3,1,*, Bing Zhang1,2,*, Yong Li4, Ping Zhang2,5, Xiaoyu Shan1, Qiuting Zhang1, Mathew Brown6, Zhenxin Li7, Joy D. Van Nostrand2, Fangqiong Ling8, Naijia Xiao2,3, Julia Vierheilig9, George F. Wells10, Yunfeng Yang1, Ye Deng11,13, Qichao Tu11, Aijie Wang13, Global Water Microbiome Consortium†, Tong Zhang12, Zhili He14,15, Jurg Keller16, Per H. Nielsen17, Pedro J. J. Alvarez18, Craig S. Criddle19, Michael Wagner9, James M. Tiedje20, Qiang He21,22,*, Thomas P. Curtis6,*, David A. Stahl23, Lisa Alvarez-Cohen24,25, Bruce E. Rittmann26, Xianghua Wen1,‡, and Jizhong Zhou2,1,25,‡

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

Microorganisms in wastewater treatment plants (WWTPs) play critical roles in water purification to protect public and environmental health, but their diversity and underlying mechanisms are poorly understood. Using a systematic global-sampling effort, we analyzed the 16S rRNA gene sequences from ~1200 activated sludge (AS) samples taken from 269 WWTPs in 23 countries on 6 continents. Our analyses revealed that the global AS bacterial communities contain ~1 billion bacterial phylotypes with a log-normal diversity distribution. Despite this high diversity, AS has a small global core bacterial community (n = 28 OTUs) that is strongly linked to AS performance. Meta-analyses with global datasets associate the AS microbiomes most closely to freshwater populations rather than human gut microbiomes. In contrast to macroorganism diversity, AS bacterial communities show no latitudinal gradient. Furthermore, their spatial turnover is scale-dependent and largely driven by stochastic processes (dispersal, drift), although deterministic factors (temperature, organic input) also are important. Our findings provide a mechanistic understanding of the global diversity and biogeography of AS bacterial communities within a theoretical ecology framework, and have important implications for microbial ecology and wastewater treatment processes.
Introduction

Microorganisms are the most phylogenetically diverse group of life on Earth, and are ubiquitous in natural and engineered environments\(^1\). They play crucial roles in the biogeochemical cycling of carbon (C), nitrogen (N), sulfur (S), phosphorus (P), and various metals. Unraveling the mechanisms generating and maintaining microbial biodiversity is key to predicting ecosystem responses to environmental changes\(^2\), as well as improving bioprocesses such as wastewater treatment and soil remediation\(^3\). With recent advances in metagenomic technologies\(^4\), microbial biodiversity and distribution are being intensively studied in a wide variety of environments\(^5\)-\(^7\), including the human gut, oceans, freshwater, air, and soil. In comparison, we are just beginning to understand the diversity and biogeography of microbial communities in engineered environments, such as activated sludge (AS) processes in wastewater treatment plants (WWTPs)\(^3,8\).

More than 300 km\(^3\) of wastewater is produced globally each year\(^9\); this equals a volume of seventh of the global flow in rivers\(^10\). About 60% of this wastewater is treated prior to release, and biological processes such as AS are widely used in WWTPs\(^9\). The AS process employs microbial flocs or granules to remove C, N, P, micropollutants (e.g., toxins, pesticides, hormones, pharmaceuticals), and pathogens\(^11\). AS is an open enrichment-culture system whose functions rely on complex and incompletely defined microbial communities. As the largest application of biotechnology in the world\(^12\), AS is a vital part of the infrastructure of modern urban societies\(^13\). Despite recent advances in understanding the microbial ecology of AS\(^14\)-\(^16\), the global picture of microbial diversity and distribution remains elusive. This information is essential to resolving controversies concerning the relative importance of stochastic versus deterministic processes in AS or related engineered ecosystems\(^3\). Such information is also important for identifying the key players of the process for future studies, and providing a basis for targeted manipulation of AS microbiomes. In addition, comparisons of microbiome data across studies are hampered by differences in experimental systems, sampling approaches, and/or analytical methods\(^4,5\).

To overcome these shortcomings, we created a Global Water Microbiome Consortium (GWMC)
to conduct a global campaign for systematically collecting and analyzing AS microbiomes. We collected AS samples from 269 WWTPs in 86 cities, 23 countries, and 6 continents (Supplementary Fig. 1a). Deep sequencing and analysis of the 16S rRNA gene using standard protocols were performed to address fundamental ecological questions, including: (i) What is the extent of global diversity of AS microbial communities? (ii) Does a core microbiome exist in AS processes across different continents? (iii) Do AS microbiomes show a latitudinal diversity gradient (LDG)? (iv) Is microbial biodiversity important to the functions of the AS process? and (v) What is the relative importance of deterministic versus stochastic factors in regulating the composition, distribution, and functions of AS microbial communities?

**Sampling representation and sequencing statistics**

The GWMC is a collaboration across more than 70 research groups from 23 countries. Unlike the Earth Microbiome Project (EMP), which employed a bottom-up strategy to solicit microbial samples\(^5\), we used a top-down approach to select WWTPs for sampling by considering their latitudes, climate zones, spatial scales, AS process type, and accessibility for sampling (Fig. 1a; see Methods for details). Altogether, we collected 1,186 AS samples using a consistent sampling protocol from global (e.g., across 6 continents), regional (e.g., individual continents), to local scales (e.g., geographic sites or individual cities).

A standardized approach was used for DNA extraction and sequencing to minimize experimental variability\(^4,17\). The V4 region of the 16S rRNA gene, a common molecular marker for bacteria and archaea\(^18\), was amplified and deeply sequenced in one laboratory using the same method for all samples. The numbers of sequences (reads) per sample ranged from 25,631 to 351,844 (Supplementary Fig. 1a), and a total of 96,148 OTUs were obtained. About 1.3% of these OTUs were from archaea, which accounted for 0.13% of the total abundance. The choice of the PCR primer pair 506F/806R (that was also used in the EMP project) is very likely to have strongly influenced this low archaeal abundance due to the much lower coverage of the primers of archaeal 16S rRNA genes compared to the bacterial counterparts (see Methods for details). 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.
Although our sequencing depths were considerably higher than those in many similar studies, rarefaction curves (Supplementary Fig. 1b, c) of AS microbial communities indicated that additional rare taxa were likely present in individual samples. Nevertheless, pooling all sequences gave a sufficient number for estimating global- and continent-level diversity of AS microbial communities (Supplementary Fig. 1d, e). All samples were rarefied to a minimum depth of 25,600 sequences for statistical analyses. This resulted in 61,448 OTUs overall, and the global OTU richness per sample was 2309±559 (Supplementary Fig. 1a). Besides richness, we also calculated other alpha diversity indices on a global and regional scale (Supplementary Fig. 1a).

Comparing our sequences with existing 16S rRNA gene databases highlights the novelty captured by the AS microbial communities. Our AS sequences match to 38.6% of Greengenes and 37.2% of SILVA 16S rRNA gene OTUs at 97% similarity. These matches accounted for 18.2% and 22.5% of the representative sequences in our datasets, respectively, indicating that the majority of AS microbial species diversity is not yet captured in full-length sequence databases; this is similar to the observations in the EMP.

Associated metadata, conforming to the Genomic Standards Consortium’s MIxS and Environmental Ontology (ENVO) Standards, were provided by plant managers and/or investigators (Supplementary Table 1; Supplementary Fig. 2). The WWTPs represent diverse geographies and a large range of climatic conditions, operation parameters, and chemical conditions across and within continents (Supplementary Fig. 2). For instance, the average influent biochemical oxygen demand (BOD), representing biodegradable organic matter, ranged from 30 to 1000 mg/L. Such a broad range of diverse parameters is critical to disentangling mechanisms of AS microbial community assembly.

**Species abundance distributions**

Species abundance distribution (SAD), which characterizes the distribution of abundances of all species within a community, is one of the few universal tools in ecology. SAD is central to
biodiversity theory and has not been rigorously tested in microbial ecology until recently\textsuperscript{24}.

Although numerous SAD models are available, log-normal and log-series have been the most successful in predicting SADs, and they are the standards for testing other models\textsuperscript{24}. While the log-series model is well supported by macroecological studies, the log-normal model is more commonly observed with microorganisms\textsuperscript{24,25}.

In this study, we tested several well-known SAD models, including log-normal, log-series, Broken-stick, and Zipf. The log-normal model explained 99\% of the variation of the AS bacterial SADs compared with 72\% for log-series, 94\% for Zipf model, and 14\% for Broken-stick (Fig. 1b; Supplementary Fig. 3a). Consistent with previous microbial studies\textsuperscript{24}, our results support that the log-normal model gives the best fit to the observed SADs.

**Extent of global microbial diversity**

One of the grand challenges in biodiversity research is determining the number of species in an ecological system\textsuperscript{26}. Given the fact that the global AS bacterial communities followed log-normal distributions, we estimated the global richness of AS bacterial communities based on two parameters\textsuperscript{25,26}. One is the total number of individuals (\(N_T\)), which was estimated as \(4 - 6 \times 10^{23}\) bacteria in AS communities, based on published data\textsuperscript{9}. The other is the quantity of the most abundant taxa (\(N_{\text{max}}\)), which can be estimated based on either our sequence data or the dominance-scaling law\textsuperscript{26}. The log-normal model predicts \(1.1(\pm 0.07) \times 10^9\) species in AS systems globally when \(N_{\text{max}}\) is estimated as \(1.2\%\) of \(N_T\) based on our sequence data. The number of species increases only slightly, to \(2.0(\pm 0.2) \times 10^9\) species, using \(N_{\text{max}} = 0.4 \times N_T^{0.93}\), the dominance-scaling law\textsuperscript{26} (Fig. 1c). AS harbored similar numbers of species for the United States (\(4.6 \times 10^8\) to \(1.1 \times 10^9\)) and China (\(3.9 \times 10^8\) to \(1.0 \times 10^9\)). Thus, the estimates of global AS bacterial richness are only about one order of magnitude lower than that of the global ocean microbiome\textsuperscript{26} (\(\sim 10^{10}\)), even though the world’s oceans represent an enormously larger ecosystem.
Global core bacterial community

Core microbiota provide information on putatively important microorganisms for system functions\textsuperscript{14}. At the global scale, occupancy-frequency and occupancy-abundance analyses revealed a hyper-dominant pattern (Supplementary Fig. 3b) in which the 866 most abundant OTUs (1.39% of the total OTU number) accounted for 50.06% of the total abundance. Similar hyper-dominance patterns were observed in other macro\textsuperscript{-27} and microbial communities\textsuperscript{28}.

A core bacterial community was determined based on the abundance and occurrence frequency of OTUs (see Methods for details). About 0.05% (28 OTUs) were identified as a global core that accounted for 12.4% of all sequences obtained (Fig. 2a; Supplementary Table 2). The majority (82%) of the core community members belonged to *Proteobacteria*, with 15 OTUs classified as β-*Proteobacteria* (Fig. 2b). Among the core community, the most abundant OTU that accounted for 1.14% of the total abundance and occurred in 85% of all samples, was 99% similar to the γ-proteobacterium *Dokdonella kunshanensis DC-3* from a WWTP\textsuperscript{29}. The second most abundant OTU (0.89% of the total abundance and occurring in 96% of all samples) belonged to *Zoogloea*, a dominant genus in AS communities\textsuperscript{15}, with *Z. ramigera* known to play important roles in the flocculation of AS\textsuperscript{30}. A *Nitrospira* OTU (OTU_6) was also identified as a core taxon, reflecting its importance for nitrite oxidation or complete ammonia oxidation in AS\textsuperscript{31,32}. In addition, OTU_7 is closely related to *Arcobacter* species, which are highly abundant in raw sewage\textsuperscript{33}, and include potential human pathogens, such as *A. cryaerophilus*, *A. butzleri*, and *A. skirrowii*\textsuperscript{34}.

Using the same criteria, we determined core communities for soil, freshwater, ocean, animal and human feces, and air at the global scale based on the EMP datasets\textsuperscript{5}. Soil, human feces, air, and freshwater microbiomes had 9, 6, 2, and 1 bacterial OTUs identified as core taxa, respectively (Supplementary Table 3). No core taxa were found for animal feces and ocean, possibly due to highly variable community compositions. Notably, the core community for AS had no overlap with the other habitats, suggesting that AS selects for a unique core community.

Latitudinal diversity pattern
Latitudinal diversity gradient (LDG), whereby species richness tends to decrease as latitude increases\textsuperscript{35}, is well documented in plant and animal ecology, but not in microbial ecology\textsuperscript{36}. Recently, several studies examined latitudinal diversity patterns in microbial communities, but the results provided no clear trend\textsuperscript{6,7,37}. In contrast to natural systems, AS operates under relatively stable and similar conditions everywhere. Thus, one might not expect AS microbial communities to exhibit LDG.

To test this hypothesis, we examined the relationship between OTU richness and latitude. OTU richness peaked at intermediate latitude, having a mean air temperature \(\sim 15^\circ\text{C}\) (Fig. 1d). As taxonomic and phylogenetic diversity were highly correlated (\(R^2 = 0.92\)), the trend was similar for phylogenetic diversity (Supplementary Fig. 1f). These results suggest that a LDG does not occur in the AS microbiomes, a result similar to what has been observed in the global ocean microbiome\textsuperscript{7}, but different from what was previously observed in some ocean\textsuperscript{37} and soil communities\textsuperscript{38}. In addition, the relationship between bacterial richness and temperature (Supplementary Fig. 1g, h) did not fit predictions from the metabolic theory of ecology\textsuperscript{39}. This theory has a very low explanatory power on bacterial richness, based on both air temperature (Supplementary Fig. 1g, \(R^2 < 0.001\)) and mixed liquid temperature (Supplementary Fig. 1h, \(R^2 = 0.03\)).

### Continental-level differences in bacterial community structure

Variations in community composition, called \(\beta\)-diversity, are key for understanding community assembly mechanisms\textsuperscript{2,40} and ecosystem functioning\textsuperscript{41}. To understand how AS bacterial community composition varied across different spatial scales, we examined taxonomic and phylogenetic diversity. First, the diversity was highest in Asia and lowest in South America (Supplementary Fig. 1a). Second, considerable variations between AS samples were observed even at the phylum level (Supplementary Fig. 3c). At the OTU level, although the taxonomic and phylogenetic structures of the AS bacterial communities were not clearly separated in two-dimensional ordinations (Supplementary Fig. 3d, e), PERMANOVA indicated that taxonomic and phylogenetic composition were significantly different (\(P < 0.001\)) between any two continents (Supplementary Fig. 3f). Third, climate and AS process type exerted significant
affects \((P = 0.001)\) on microbial community structure, but these were overwhelmed by continental geographical separation (Supplementary Fig. 3g). For example, bacterial communities of the same climate type ‘Cfa’ in North America and Asia were separated by their continental origins rather than being clustered together (Supplementary Fig. 3h, i). The AS bacterial communities had relatively higher similarity to those of freshwater and soil than to human feces (Fig. 3a). However, AS harbored a unique microbiome distinctly different from all other habitats, including soil, ocean, freshwater, air, and human and animal feces (Supplementary Table 4).

A Bayesian approach\(^{42}\) was employed to identify potential sources of AS bacterial communities at the genus level. The most dominant potential source was found to be freshwater, attributing on average 46% of genera, followed by soil (17% on average) and ocean (12% on average) (Fig. 3b). This could be because the environmental characteristics are more similar between an AS bioreactor and freshwater than to the others. Also, AS and freshwater have potentially high immigration events through connected water systems, such as wastewater being input and discharged to rivers after treatment.

**Scale-dependent distance-decay patterns**

Another fundamental pattern in ecology is the distance-decay relationship (DDR)\(^{23,43}\), in which community similarity decreases as geographic distance increases. Consistent with results in other domains\(^{40}\), we hypothesized that (i) the slope of the DDR curve would vary over local, regional, and global scales, and (ii) the spatial turnover rates of AS microbial communities would be lower than those observed in natural habitats, especially for non-flowing ecosystems, such as soils\(^{44}\).

To test these hypotheses, we first measured the DDRs of the AS bacterial communities by linear regression between log-transformed community similarity and spatial distance. Supporting our first hypothesis, significant negative DDRs \((P < 0.001)\) were observed across all scales based on taxonomic diversity (slope = -0.06 for Sorensen and -0.08 for Bray-Curtis) and phylogenetic diversity (slope = -0.04 for unweighted Unifrac, and -0.02 for weighted Unifrac) (Fig. 4a, 4b).
Supplementary Fig. 4a). The slopes of DDRs depended significantly on spatial scale. The DDR slopes across cities within a continent (-0.13 ~ -0.15 for taxonomic similarity indices; -0.03 ~ -0.085 for phylogenetic similarity indices) were significantly ($P = 0.001$) steeper (> 2 times) than the overall slopes for all similarity metrics (Supplementary Fig. 4a). Countering our second hypothesis, the overall spatial turnover rates of the AS communities were similar to those found in non-flowing natural habitats such as soils and sediments.

**Relationships between the community structure and AS functions**

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

To assess functions, we calculated the removal rates of organic matter (BOD, COD), total phosphorus (TP), total nitrogen (TN), and ammonium nitrogen (all in g chemical per MLSS per day, where MLSS is mixed liquor suspended solids relating to microbial biomass). Partial Mantel tests revealed that the distance-corrected changes of AS-community composition were significantly correlated with all measured removal rates ($P < 0.032$), except for the ammonium-nitrogen removal rate ($P > 0.18$) (Supplementary Fig. 5a). Consistent with our expectation, the AS community composition was significantly correlated with the TP removal rate for samples from enhanced biological phosphorus removal (EBPR) plants, but not for non-EBPR plants (Supplementary Fig. 5a), as P removal processes in non-EBPR plants are predominantly chemical over biological. The nitrifying microbial community, which included *Nitrospira* and *Nitrosomonas* OTUs, showed a closer correlation with the ammonium nitrogen removal rate than did the whole community (Supplementary Fig. 5a; $P$ of Bray-Curtis distance =0.04). Further analysis revealed positive correlation of *Nitrospira* (Spearman’s $\rho = 0.41$, adjusted $P < 0.001$) and *Nitrosomonas* (Spearman’s $\rho = 0.23$, adjusted $P < 0.001$) abundance with the percent of ammonium-nitrogen removal (% of influent concentration), but not with the ammonium-nitrogen
removal rate (Supplementary Fig. 5b), which implies a potential trade-off between the quality of the ammonium nitrogen removal (removal percentage) and the efficiency of the removal (removal rate). In fact, *Nitrospira* was the top genus correlating with the percent of ammonium-nitrogen removal, further highlighting its role in nitrite oxidation or complete ammonia oxidation in AS. Of the 28 global core OTUs, 27 were significantly correlated (adjusted $P < 0.05$) with at least one of the five functions examined. Most of the correlations (81%) were positive (Fig. 2c). Also, about 80% of the non-core OTUs showed significant correlations (adjusted $P < 0.05$) with at least one function, and 40% of these correlations were positive (Supplementary Fig. 5c). All of these results indicated that the structure of the AS bacterial communities, particularly the dominant populations, is critical to maintaining AS functions.

**Stochastic community assembly**

Various global structural and functional patterns appear in the AS communities, but what are the underlying mechanisms? Since WWTPs are well controlled engineered ecosystems, we hypothesized that the AS community has a deterministic nature. To test this hypothesis, we calculated the null model-based stochastic ratios (ST)\(^{44}\) with taxonomic (Bray-Curtis/Sorensen) and phylogenetic (weighted and unweighted Unifrac) metrics. The average ST ratios based on these four metrics all were higher than 0.75 (Fig. 4b), meaning that stochastic processes could play more important roles in influencing community composition than deterministic processes, at least partially contradicting our hypothesis.

To further identify the relative importance of various factors contributing to spatial turnover of the AS 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, 4d, Supplementary Fig. 4b). 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 (Fig. 4d, Supplementary Fig. 4b). These results are consistent with the null model analysis indicating that the community structure is controlled primarily by stochastic processes.

The effects of geographic distance on community composition varied substantially with spatial scale: 28% within cities, 2% within continents, and almost nothing across continents (Fig. 4d). In contrast, environmental variables exerted influence in the opposite direction, with 2% within cities, 14% within continents, and 24% across continents (Fig. 4d). Similar trends were observed with other similarity metrics (Supplementary Fig. 4b). These results suggest that, as the spatial scale increases, geographic effects decrease, but environmental effects increase. This could be explained by the combined influences of ecological drift and dispersal barrier at large scales (e.g., between continents)\(^{43}\), as well as unmeasured environmental variables that are spatially auto-correlated at small scales (e.g., within a city) (Supplementary Fig. 4b)\(^{40}\).

**Environmental drivers of community composition**

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

More in-depth analysis by structural equation modeling (SEM) revealed direct and indirect effects of the environmental drivers (Fig. 5b). For example, consistent with Mantel test, temperature had the strongest direct effects on PC1 representing the community structure (standardized path coefficient, \(\beta = 0.50, P < 0.001\)). It also had weak negative impacts on species richness \((\beta = -0.14, P < 0.001)\). Furthermore, various biotic (e.g., PC1) and abiotic
factors (e.g., food-to-microorganisms ratio [F/M] indicating the level of organic matter supplied to microorganisms, dissolved oxygen, and SRT) directly affected BOD-removal rates (Fig. 5b). The influent BOD likely has an impact on bacterial composition through its effect on the F/M ratio ($\beta = 0.31, P < 0.001$), which is inversely related to the SRT. The influent BOD level 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 level ($R^2 = 0.19, P < 0.0001$; Fig. 5c). All of these results are consistent with resource-competition theory\textsuperscript{48}, which predicts that high species diversity occurs with low to intermediate supply of resources, but fast-growing r-strategists can outcompete efficient-scavenging K-strategists at high resource levels\textsuperscript{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. 5d, Supplementary Fig. 6a). To control for spatial autocorrelation, we further excluded nearest geographical neighbors from the training set. Models of temperature continued to have higher accuracy at different excluded-radius scales (Supplementary Fig. 6a). For example, the America-fitted model of temperature, i.e., a model trained solely by North America and South America samples, was able to capture variations in the temperatures of Asia samples (cross-validated $R^2 = 0.47$) (Fig. 5d). The random forest model also revealed the most important OTUs for predicting temperature (Supplementary Fig. 6b). These results support the finding that temperature is the major environmental variable shaping the AS bacterial compositions at the global scale, although it only has weak effect on species richness (Fig. 5b).

**Conclusions and future perspectives**

In summary, determining microbial diversity, distributions, and their underlying functions are critical challenges in microbial ecology, particularly at a global scale. Through well-coordinated international efforts, we systematically examined global diversity and biogeography of AS
bacterial communities within the context of theoretical ecology frameworks. Our findings greatly enhance the understanding of microbial ecology in AS, and set 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 AS systems are globally inhabited by ~ $10^9$ different bacterial species. In contrast, only about $10^4$ species have been cultivated and studied in detail\textsuperscript{26}. If we assume that all cultivated species are present in AS, potentially 99.999% of AS microbial taxa remain uncultured. Although more and more microorganisms have been genomically characterized, exploring the physiological attributes, which requires cultivation, represents a formidable task for future microbiologists and process engineers\textsuperscript{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 AS, 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 AS is a managed ecosystem, its bacterial composition appears to be driven mainly by stochastic processes, such as dispersal and drift, which contradicts conventional wisdom. However, deterministic processes, e.g., temperature, SRT, and organic C inputs, play important roles in regulating the structure of the AS community. This finding could be important for developing operating strategies to maintain biodiversity that promotes stable system performance. Perhaps one could overcome dispersal limitation by establishing new, or repopulating degraded or failed WWTPs, by inoculating them with AS from functioning WWTPs, which is a common practice in environmental engineering. Alternately, one could alternate organic C loadings and/or operational conditions to manipulate the AS community’s structure to select for the microorganisms having the desired functions.

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


Acknowledgements

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

Author contributions


Global Water Microbiome Consortium

Members besides the authors listed on the first page:

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

\textsuperscript{1}Environmental Microbiology and Biotechnology Laboratory, Engineering School of Environmental & Natural Resources, Engineering Faculty, Universidad del Valle–Sede Meléndez, Cali, Colombia; \textsuperscript{2}Advanced Water Management Centre, The University of Queensland, Brisbane, QLD, Australia; \textsuperscript{3}Department of Environmental Science, Policy, and Management, University of California, Berkeley, USA; \textsuperscript{4}Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, USA; \textsuperscript{5}Universidade Federal de Minas Gerais, Departamento de Engenharia Sanitária e Ambiental, Brazil;

\textsuperscript{6}Department of Environmental Health Sciences, The University of Michigan, Ann Arbor, MI, USA; \textsuperscript{7}Hampton Roads Sanitation District (HRSD), Virginia Beach, VA, USA; \textsuperscript{8}Microbial Ecology Laboratory, Microbial Biochemistry and Genomics Department, Biological Research Institute “Clemente Estable”, Montevideo, Uruguay; \textsuperscript{9}Institute of Water and Wastewater Technology, Durban University of Technology, South Africa; \textsuperscript{10}Aix-Marseille University CNRS IRD, MIO UM110 Mediterranean Institute of Oceanography, Marseille, France; \textsuperscript{11}Escuela de Ingeniería Bioquímica, Pontificia Universidad Católica de Valparaíso, Valparaíso, Chile;

\textsuperscript{12}Department of Civil and Environmental Engineering, The University of Tennessee, Knoxville, TN, USA; \textsuperscript{13}Department of Computer Science, Tsinghua University, Beijing, China;

\textsuperscript{14}Bioinformatics Division, Tsinghua National Laboratory of Information Science and Technology, Tsinghua University, Beijing, China; \textsuperscript{15}Microbial Community Engineering Laboratory, Department of Civil Engineering and Applied Mechanics, McGill University, Montreal, Canada; \textsuperscript{16}Department of Chemical and Biological Engineering, Northwestern University, Evanston, IL, USA; \textsuperscript{17}Department of Civil and Environmental Engineering, Northeastern University, Boston, MA, USA; \textsuperscript{18}Vatten & Miljö i Väst AB (VIVAB), Falkenberg, Sweden; \textsuperscript{19}Institute for Environmental Genomics, Department of Microbiology and Plant Biology, and School of Civil Engineering and Environmental Sciences, University of Oklahoma, Norman, OK, USA; \textsuperscript{20}Norman Water Reclamation Facility, Norman, OK, USA; \textsuperscript{21}Golden Heart Utilities, Fairbanks, AK, USA; \textsuperscript{22}Karlsruhe Institute of Technology, Engler-Bunte-Institut, Water Chemistry and Water Technology, Germany; \textsuperscript{23}Department of Civil and Environmental Engineering, University of Missouri, Columbia, MO, USA; \textsuperscript{24}Department of Building, Civil and Environmental Engineering, Concordia University, Montreal, Canada; \textsuperscript{25}Department of Environmental Microbiology, Eawag, Dübendorf, Switzerland; \textsuperscript{26}Water Resources Engineering, Faculty of Engineering, Lund University, Lund, Sweden; \textsuperscript{27}Luxembourg Centre for Systems Biomedicine, University of Luxembourg, Belvaux, Luxembourg; \textsuperscript{28}Department of Civil and Environmental Engineering, Yonsei University, Seoul, South Korea; \textsuperscript{29}Advanced Environmental Biotechnology Centre, Nanyang Environment & Water Research Institute, Nanyang Technological University, Singapore; \textsuperscript{30}Department of Civil and Environmental Engineering, Rice University, Houston, TX, USA; \textsuperscript{31}Department of Civil Engineering, University of Nebraska–Lincoln, NE, USA; \textsuperscript{32}School of Civil and Environmental Engineering, Nanyang Technological University, Singapore; \textsuperscript{33}Department of Civil and Environmental Engineering, Stanford University, Stanford, CA, USA; \textsuperscript{34}Plant Biotechnology Program, Federal University of Rio de Janeiro, UFRJ, Brazil; \textsuperscript{35}Federal University of Ceará, UFC, Brazil; \textsuperscript{36}University of Tennessee, Center for Environmental Biotechnology, Knoxville, TN, USA; \textsuperscript{37}Department of Civil Engineering, Construction Management and Environmental Engineering, Northern Arizona University, Flagstaff, AZ, USA;
University, Flagstaff, AZ, USA; 38UCIBIO, REQUIMTE, Departamento de Química, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa, Caparica, Portugal; 39CSIRO Land and Water, Ecosciences Precinct, QLD, Australia; 40Biodesign Swette Center for Environmental Biotechnology, Arizona State University, Tempe, AZ, USA; 41Departamento de Química, Universidade de São Paulo, Faculdade de Filosofia Ciências e Letras de Ribeirão Preto – FFCLRP, Ribeirão Preto, SP, Brazil; 42Department of Civil, Construction, and Environmental Engineering, North Carolina State University, NC, USA; 43Grupo de Investigación en Procesos Anaerobios, Instituto de Ingeniería, Universidad Nacional Autónoma de México, México, DF, Mexico; 44CNR-IRSA, National Research Council, Water Research Institute, Rome, Italy; 45Infrastructure and Environment Research Division, School of Engineering, University of Glasgow, Glasgow, UK; 46Tryon Creek & Columbia Blvd. Wastewater Treatment Plants, Bureau of Environmental Services, City of Portland, OR, USA; 47Scion Research, Christchurch, New Zealand; 48School of Engineering, University of Guelph, Guelph, Ontario, Canada; 49Department of Environmental Engineering, National Cheng Kung University, Tainan City, Taiwan, ROC; 50State Key Laboratory of Applied Microbiology Southern China, Guangdong Provincial Key Laboratory of Microbial Culture Collection and Application, Guangdong Institute of Microbiology, Guangzhou, Guangdong, China; 51Department of Civil and Environmental Engineering, University of Hawaii at Manoa, Honolulu, HI, USA; 52State Key Laboratory of Environmental Aquatic Chemistry, Research Center for Eco-Environmental Sciences, Chinese Academy of Sciences, Beijing, China; 53State Key Joint Laboratory of Environment Simulation and Pollution Control, School of Environment, Tsinghua University, Beijing, China; 54Department of Civil Engineering, University of Arkansas, Fayetteville, AR, USA; 55Earth and Environmental Sciences, Lawrence Berkeley National Laboratory, Berkeley, CA, USA; 56School of Engineering, Newcastle University, Newcastle upon Tyne, UK; 57Institute for a Secure and Sustainable Environment, The University of Tennessee, Knoxville, TN, USA; 58Environmental Microbiomics Research Center, School of Environmental Science and Engineering, Sun Yat-Sen University, Guangzhou, China; 59Guangdong Provincial Key Laboratory of Environmental Pollution Control and Remediation Technology, Guangzhou, China; 60Consolidated Core Laboratory, University of Oklahoma, Norman, Oklahoma, USA.

Competing interests

The authors declare no competing financial interests.

Materials & Correspondence

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

Global sampling and meta-data collection

The Global Water Microbiome Consortium (GWMC) was initiated in May 2014 as a platform to facilitate international collaboration and communication on research and education for global water microbiome studies (http://gwmc.ou.edu/). 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 municipal WWTP microbiomes. Towards this goal, WWTPs were selected based on the following criteria:

(i) Continental-level geographic locations. Samples were obtained from all continents except for Antarctica, but with special focus on North America, Asia, and Europe (Fig. 1a). Because of the low accessibility, WWTPs in Africa and South America were under-represented. (ii) Latitude. To address questions related to latitudinal diversity gradient (LDG), WWTPs were intensively sampled in North America along the East and West Coasts, and Highway 35, as well as Highway 40 (from East to West) (Fig. 1a), in Asia, Europe, and Australia. The WWTPs sampled spanned latitudes from 43.6°S to 64.8°N. (iii) Climate zones. Since climate could have significant impacts on microbial communities, the samples covered 17 different climate types (Supplementary Fig. 7). To be able to distinguish independent effects of continents versus climate zones, we increased sampling efforts for climate zones that were present in multiple continents, such as Humid Subtropical Climate. (iv) Scales. The samples were collected from very broad spatial scales: global (across 6 continents), regional (e.g., individual continents or climate zones), and local (e.g., individual cities). Within some cities, multiple WWTPs and multiple samples per WWTP were collected; (v) Wastewater treatment process types. To be able to address the relationship of structure to function for AS, we sampled the aerobic zone of conventional plug flow, oxidation ditch, sequential batch reactors, anaerobic/anoxic/oxic (A²O), and other AS process types.

A top-down design was implemented with unified protocols for sampling, sample preservation, metadata collection, DNA extraction, sequencing, and sequence analysis, to minimize potential
experimental variations\textsuperscript{4,17,51,52}. The detailed sampling and metadata collection methods and protocols are available at the GWMC web site (http://gwmc.ou.edu/protocols/view/11).

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

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

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. Raw metadata were processed as one metadata table (Supplementary Table 1) and classified into three categories: geological variables, plant operation and monitoring variables, and sample properties. The geological variables included latitude and longitude; ambient climate variables such as climate type, mean annual temperature (MAT), and precipitation; and population size and gross domestic product (GDP) for the city where the WWTP was located.
Climate type was determined by the Köppen-Geiger climate classification. GDP and population data were derived from the Brookings analysis of Global Metro Monitor. Variables related to plant design and operation include plant age, design capacity, actual flow rate, volume of aeration tanks, hydraulic retention time (HRT) and solids retention time (SRT). The AS process type, aerator type, and coupling with N removal processes (nitrification and denitrification) in the WWTP was also provided by the plant managers as possible. Plant monitoring variables include influent and effluent biochemical oxygen demand (BOD) and chemical oxygen demand (COD) representing organic carbon (C) level, total nitrogen (TN) and total phosphorus (TP) representing nutrient level, ammonium N, as well as the food to microorganism (F/M) ratio, indicating the average organic C loading to microorganisms. For sample properties, most plant managers provided the yearly average value of mixed liquor suspended solids (MLSS), indicating the concentration of biomass in the AS, dissolved oxygen (DO), pH, and mixed liquid temperature; some provided the measured values when sampling.

AS performance was calculated as the specific removal rates (g per g biomass per day) of organic C (BOD and COD), nutrients (TN and TP) and ammonium nitrogen (NH$_4$-N):

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

**DNA Extraction**

To minimize the variations associated with sample processing, identical protocols were used in DNA extraction and 16S rRNA gene sequencing. All samples from China and Japan were shipped to Dr. Xianghua Wen’s Laboratory at Tsinghua University for DNA extraction. All other samples, including samples from Europe collected by Dr. Thomas Curtis at Newcastle University, were shipped to Dr. Jizhong Zhou’s Laboratory at University of Oklahoma (OU) for DNA extraction. Due to the tight restriction of sample shipment in South Africa, Mexico, Chile, Uruguay, and Brazil, the DNA was extracted by GWMC members in these countries. DNA was extracted from sludge samples using MoBio PowerSoil DNA isolation kit. For each sample, a pellet from 3 mL mixed liquor was used. In addition to the manufacture protocol, we always placed exactly 12 bead tubes on the vortex evenly and vortex at maximum speed for 10 min to
minimize the lysis efficiency difference between samples. All DNA samples were processed at OU for sequencing.

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

**16S rRNA gene sequencing and sequence processing**

The V4 region of the 16S rRNA gene was amplified and sequenced using standardized protocols with the phasing amplicon sequencing (PAS) approach as described previously and the primers 515F (GTGCCAGCMGCGCGGTAA) and 806R (GGACTACHVGGGTWTCTAAT) of the Earth Microbiome Project. *In silico* primer coverage analysis using SILVA TestPrime 1.0 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, the 16S rRNA gene fragments were amplified from community DNAs (10 ng) with two-step PCR using lower numbers of amplification cycles (10 and 20 cycles for the 1st and 2nd step, respectively). The two-step PAS approach offers several advantages: lower amplification biases, better sequence-read quality, higher effective sequence read numbers and length, and lower sequencing errors. All samples were sequenced using the same MiSeq instrument at the Institute for Environmental Genomics, OU. Generally, around 400 samples were combined together for each round of MiSeq sequencing. Since the numbers of sequence reads varied substantially from sample to sample, most samples were sequenced more than once (e.g., 19% twice; 33%, three times; 43%, > 3 times) to meet the target number of about 30K sequencing reads per sample, as determined in our previous analysis.

Raw sequence data were processed as previously described, except for OTU generation by UPARSE at the 97% similarity threshold, resulting in 96,148 OTUs. We define operational
taxonomic units (OTUs) (based on 97% sequence similarity) for bacterial and archaeal phylotypes. Although there is potential misconnection between OTUs and microbial species, we use this popular definition for simplicity, and it also allows us to compare with previous studies of other systems. The representative sequences were aligned using Clustal Omega v1.2.2 for constructing the phylogenetic tree by FastTree2 v2.1.10. The OTUs were taxonomically annotated with RDP Classifier using 16S rRNA gene training set 16 with a confidence cutoff of 50%. After removal of the global singletons, the sequence number in each sample was rarefied to the same depth (25,600 sequences per sample) in subsequent comparative analyses. The sequencing depth used for our analysis is ~2-5 times higher than in some similar global survey studies.

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

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 AS samples were compared against the representative set (97% similarity level) of full-length sequences from Greengenes 13.8 (released on August 2013) and the non-eukaryotic fraction of Silva 132 databases (released on December 2017). We used the open-source sequence search tool USEARCH10 in global alignment search mode, and we required 97% similarity across the query sequence.

Species abundance distribution (SAD) fitting

We compared the SAD of each sample, based on the rank-abundance distribution, with predictions from log-normal, log-series, Broken-stick, and Zipf models. By comparing (rank-for-rank) the observed and predicted SADs using regression analysis, we could directly infer the percentages of variations in abundance among species explained by each model using the same code, developed by Shoemaker et al.
Estimation of global bacterial diversity of WWTPs

We used the methods described in Curtis et al.25 and Locey and Lennon26 to predict global bacterial richness ($S_T$) using the log-normal model. The log-normal prediction of $S_T$ is based on the total abundance ($N_T$), the abundance of the most abundant species ($N_{max}$), and the assumption that the rarest species is a singleton, $N_{min} = 1$. In communities with $N_T$ individuals, the richness can be estimated by:

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

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

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

We used published data to estimate the total microbial abundance in WWTPs as follows. Empirical records compiled from a variety of sources, for example, AQUASTAT67 and Sato et al 201368, suggest that about 330 km$^3$ year$^{-1}$ of municipal wastewater are produced globally, of which 60% is treated9. Assuming that they are all treated in WWTPs, then about 0.54 km$^3$ municipal wastewater are treated by WWTPs globally per day. The total effective volume of aerobic tanks of WWTPs can be estimated by:

$$V = Q \times HRT \quad \text{(iii)}$$

where Q is the influent flow rate (m$^3$ day$^{-1}$) and HRT is the hydraulic retention time (day) of the aerobic tank. Our dataset indicates that the average HRT of aerobic tanks is 9.8 (± 0.3 s.e.) hours. Thus, the total effective volume is estimated as 0.22 (± 0.007) km$^3$. The total cells in AS are about 2.3 (± 0.4)$\times 10^9$ (ml$^{-1}$)69; thus, $N_T$ (global AS bacterial abundance) is about 4.0 - 6.1 $\times 10^{23}$. 

27
We then estimated \( N_{\text{max}} \) based on the ratio of \( N_{\text{max}} \) to \( N_T \) of our sequencing data, i.e., the relative abundance of the most abundant OTU, or using scaling law\(^{26} \). The knowledge of \( N_T \), \( N_{\text{max}} \), and \( N_{\text{min}} \) allows equation (ii) to be solved numerically for the parameter \( \alpha \) and, subsequently, for \( S_T \) using equation (i).

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 cells about \( 2.3 (\pm 0.4) \times 10^{19} \). \( N_T \) of WWTPs in US and China were estimated based on their published data of treating amount\(^{70,71} \). \( N_{\text{max}} \) was further estimated based on our 16S rRNA gene sequencing data or using a scaling law\(^{26} \). The total bacterial richness estimates of individual human gut, individual cow rumen, global ocean and Earth were taken from Locey and Lennon\(^{26} \).

**Core community determination**

A global-scale core microbial community was determined based on multiple reported measures. First, “overall abundant OTUs” were filtered out according to mean relative abundance across all samples (MRA)\(^{72} \). Previous studies used different criteria (e.g., MRA > 1\%\(^{73,74} \) or 0.1\%\(^{75,76} \)) without any objective or standard rule. Thus, we selected all top 0.1\% OTUs (62) as overall abundant OTUs. Their MRA was higher than 0.2\%, within the range of reported criteria. Second, “ubiquitous OTUs” were defined as OTUs with occurrence frequency in more than 80\% of all samples\(^{77} \). Finally, “frequently abundant OTUs” were selected based on their relative abundances with a sample (RA). In each sample, the OTUs were defined as abundant when they had a higher RA 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 samples, which is stricter than the reported criterion (10 in 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). 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 AS OTUs were repicked using closed reference against Greengenes 13.8. This OTU table was then merged with the EMP OTU table. To give relatively equal representation of samples across environments, we further collapsed our AS samples at the plant level by summing the abundance of each OTU across samples of the same plant, resulting in 269 AS samples. Our AS samples and the EMP samples from freshwater (including that from freshwater and freshwater biofilm), ocean (including that from sea water and biofilm), animal feces, human feces, 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 (NMS) analysis was performed using the Bray-Curtis dissimilarity matrix.

The proportion of each AS microbiota attributable to freshwater, soil, ocean, animal and human feces, and air at the genus level were estimated using SourceTracker\(^4^2\), which was run through QIIME with default settings using AS 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 modeling.

**Diversity analyses: \(\alpha\)- and \(\beta\)-diversity and correlation with environment**

Richness and Faith’s index were used to measure taxonomic and phylogenetic \(\alpha\)-diversity, respectively, and they were computed using the *Picante* R package\(^7^8\). Other taxonomic \(\alpha\)-
diversity indices, including Shannon index, Simpson index and Pielou’s evenness, were calculated using the *vegan* R package\(^{79}\).

Bray-Curtis (abundance-based) and Sorensen (incidence-based) distances were calculated to represent the taxonomic \(\beta\)-diversity using the *vegan* R package\(^{79}\). The weighted (abundance-based) and unweighted UniFrac (incidence-based) distance\(^{80}\) were calculated to represent the phylogenetic \(\beta\)-diversity using the *GUniFrac* R package\(^{81}\). 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\(^{79}\).

PERMANOVA was applied to assess the difference of community composition among continents, climate types, and activated sludge process types using the *vegan* R package\(^{79}\). In PERMANOVA, climate types were defined at main climate group level, which includes 5 groups: A (tropical), B (arid), C (temperate), D (cold), and E (polar)\(^{53}\). The AS process types were classified into 9 general groups: complete mix, conventional plug flow, sequential batch reactors (SBR), anaerobic/anoxic/oxic (A\(^2\)O), anoxic/oxic (AO), oxidation ditch, contact stabilization, pure oxygen and extended aeration.

**Distance Decay relationships**

The rate of the distance-decay relationship (DDR) was calculated as the slope of a linear least squares regression on the relationship between ln-transformed geographic distance versus ln-transformed bacterial community composition similarity. We used matrix permutation tests to examine the statistical significance of the distance-decay slope\(^{40}\). The samples were permuted 999 times, and the observed slope was compared with the distribution of values in the permuted datasets. We also tested whether the slopes of the distance-decay curve at the three spatial scales (0 to 100 km; 100 to 5,000 km; and 5,000 to 25,000 km) were significantly different from the slope of the overall distance-decay curve, using matrix permutations to compare the observed difference between slopes within the three spatial scales with the overall distance-decay slope to that over 999 permutations.
Estimating stochasticity of community assembly

We assessed community-assembly stochasticity with a null-model-based index. The Stochasticity ratio (ST) was described previously. 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 ST using taxonomic and phylogenetic metrics. Whether using the Bray-Curtis (abundance-weighted) or Sorensen (unweighted) model, the ST was calculated based on typical null-model algorithms for taxonomic metrics. When using weighted and unweighted Unifrac, the ST was calculated based on typical null-model algorithms for phylogenetic metrics. Samples within each city were considered sharing the same regional species pool in null model algorithms.

Partitioning the environment and distance effect

To give a quantification of relative contribution of the environment effect versus the distance effect on β-diversity, we performed a variation partition analysis (VPA) based on multiple regression on matrices (MRM). We used a modified MRM approach as described previously. Briefly, we first selected a non-redundant environmental variable set. The final set included temperature, precipitation, design capacity, SRT, DO, 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. Geographic distance and microbial community distance were ln-transformed. A Euclidean distance matrix was calculated for each environmental variable. To reduce the effect of spurious relationships between variables, we first ran the MRM test with all the variables in the non-redundant environmental variable set, removed the non-significant variables from this initial MRM test, and then reran the test. The significance of the partial regression was tested by matrix permutation for 999 times. 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 ($R^2_τ$) were used to compute the four components of variations as described elsewhere.
pure environmental variation = $R^2_T - R^2_G$; (ii) pure geographical distance = $R^2_T - R^2_E$; (iii)
spatially structured environmental variation = $R^2_G + R^2_E - R^2_T$; and (iv) unexplained variation
= $1 - R^2_T$.

**Structural equation model (SEM)**

SEM was used to explore the direct and indirect relationships among environmental variables,
bacterial communities, and AS function. The community composition was represented by the
first principal component (PC1) of Principal coordinate analysis (PCoA) based on Bray-Curtis
distance. We first considered a full model that included all reasonable pathways, and then we
sequentially eliminated non-significant pathways until we attained the final model whose
pathways all were significant. To capture the quadratic correlation of SRT to diversity and BOD
removal, we constructed a composite variable of ‘SRT effect’ as a linear combination of SRT
and the square of SRT (SRT.SQ). We used a $\chi^2$ 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.

**Random Forest models**

We applied a machine-learning model, random forest, to examine the strengths of the
associations between environmental variable and compositional data, using the randomForest R
package. We used OTUs as predictors and environmental variable as response data. To
correct the potential spatial autocorrelation, we used OTU data at the plant level, by averaging
the relative abundance of each OTU across samples of the same plant. OTUs which were
detected in at least 20% of all the plants and in all continents were used for modelling. We
allowed a baseline model to learn using the full data-set for training, and subsequently, we
trained new random forests for each plant with customized training sets that excluded plants
within a defined radius of the target plant. The size of this radius ranged from 0 to 5000 km. To
delineate the model prediction strength, the cross-validated $R^2$ was calculated as

$$1 - \frac{\sum(y_i - \hat{y}_i)^2}{\sum(y_i - \bar{y})^2},$$

where $y_i$ is the value of the parameter for sample $i$, $\hat{y}_i$ is the prediction for that same sample
(obtained by held-out cross-validation), and $\bar{y}_i$ is the overall mean (the summation runs over all the samples).

### Data and code availability

The sample metadata are available in Supplementary Table 1. Raw sequence data will be made public later. No custom algorithms or software were used to generate and analyze data. R codes on the statistical analyses are available from the corresponding authors upon reasonable request.

### References of Methods


1077 72 Hanski, I. Dynamics of regional distribution: the core and satellite species hypothesis. Oikos, 210-221 (1982).
1081 74 Lawson, C. E. et al. Rare taxa have potential to make metabolic contributions in enhanced biological phosphorus removal ecosystems. Environmental microbiology 17, 4979-4993 (2015).
1113 89 Rosseel, Y. lavaan: an R package for structural equation modeling and more Version 0.4-9 (BETA). (Ghent University, 2011).


Fig. 1. The Global Water Microbiome Consortium captures microbial diversity of globally distributed wastewater treatment plants (WWTPs).  

(a) Geographical distribution of 269 WWTPs where activated sludge (AS) samples and environmental data were collected.  

(b) Predicting species abundance distribution (SAD) of AS bacterial communities.  The grey line represents a SAD that was randomly chosen from our data.  Each model was fit to the observed SAD (see Methods).  Supplementary Fig.3a shows the variations of the SADs explained by each model across all 1186 AS communities, indicating the best performance of the log-normal model.  

(c) Estimation of AS microbial richness of WWTPs.  Microbial species are defined as OTUs at 97% sequence similarity threshold.  The microbial richness-abundance scaling relationship (dashed grey line with pink hull as 95% prediction interval), and the grey dots representing richness estimates from other systems were derived from Locey and Lennon\textsuperscript{26}.  

Richness was predicted from the log-normal model using $N_T$ estimated from published data, and $N_{\text{max}}$ inferred from our sequencing data (filled circle) or $N_{\text{max}}$ predicted from the dominance-scaling law\textsuperscript{26} (hollow circles).  ‘WWTP’ indicates one WWTP, as do ‘Human gut’ and ‘Cow rumen’.  

(d) Latitudinal distribution of AS bacterial diversity, plotting OTU richness against the

**Global Water Microbiome Consortium**

- **North America**: 3 countries, 36 cities, 127 plants, 616 samples
- **Europe**: 9 countries, 14 cities, 34 plants, 107 samples
- **Asia**: 4 countries, 19 cities, 73 plants, 248 samples
- **South America**: 4 countries, 6 cities, 11 plants, 80 samples
- **Africa**: 1 country, 3 cities, 6 plants, 36 samples
- **Australasia**: 2 countries, 8 cities, 18 plants, 99 samples
absolute latitude of sampling locations shows the peak of richness at intermediate latitude. The
color gradient denotes the annual mean air temperature. Shapes of symbols denotes whether a
sample originated from Northern (circle) or Southern Hemisphere (square).
Fig. 2. Abundance, composition and functional importance of the global core OTUs in AS.

(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 global scale, which accounted for 12.4% of all 16S rRNA gene sequences. (b) The taxonomic composition of the global core OTUs on phylum and class level. (c) AS functions were calculated as the removal rate of organic carbon (BOD removal, COD removal), nutrients (TN and TP removal) and ammonia nitrogen (NH$_4$N removal) (g chemical per g MLSS per day). The color gradient on the right indicates Spearman’s correlation coefficients, and the asterisks denote the significance levels: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. In the correlation analysis, all OTUs detected in at least 20% of samples were included, and $P$-values were adjusted for multiple testing using the Benjamini and Hochberg false discovery rate (FDR) controlling procedure. Only global core OTUs were shown, with their relative abundance indicated on the left of the heatmap.
Fig. 3. Comparing bacterial community compositions across continents and with other habitats. (a) Nonmetric Multidimensional Scaling analysis (NMDS) showing that AS of WWTPs harbored a unique microbiome as compared with other habitats. For comparison, we merged our OTU table with that released by EMP⁴, which contained thousands of bacterial communities from various habitats such as soil, ocean, freshwater, air, human and animal feces, but not AS from WWTPs (see Methods for details). Bray-Curtis distance was calculated to represent the dissimilarity in bacterial community compositions. (b) Percentage of AS bacterial genera attributable to air, animal and human feces, freshwater, ocean and soil, as determined by SourceTracker.
Fig. 4. **Spatial turnover of the AS bacterial communities.** (a) Distance-decay relationships (DDRs) based on Bray-Curtis similarity. The black line denotes the least-squares linear regression across all spatial scales. The colored lines denote separate regressions within each of the three spatial scales: within cities, within continents (across cities within a continent), and intercontinental (across continents). The slopes of all lines (except the purple line) are significantly less than zero. DDRs based on other similarity metrics are listed in Supplementary Fig. 4a. (b) Ecological stochasticity in bacterial community assembly estimated by stochasticity ratio (ST), which is calculated for each pair of samples 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 (VPA) showing the relative contributions of geographic distance (Geo) and environmental variables (ENV) to the variations of community compositions of all samples based on Bray-Curtis distance. (d) VPA analyses at different spatial scales. The results based on other metrics are presented in Supplementary Fig. 4b.
Fig. 5. Environmental drivers of the AS community composition. (a) Pairwise comparisons of environmental factors are shown with a color gradient denoting Spearman’s correlation coefficients. Bacterial taxonomic community composition (Bray-Curtis distance) was related to each environmental factor by partial (geographic distance–corrected) Mantel tests. Edge width corresponds to the Mantel’s rm statistic, and edge color denotes the statistical significance based on 999 permutations. SRT is the solids retention time, representing the average time the AS solids are in the system. F/M: ratio is the food to microorganisms, representing the level of organic matter supplied to microorganisms. (b) A structural equation model (SEM) shows direct and indirect relationships among environmental variables, community composition, and WWTP functioning. To capture the quadratic correlation of solids retention time (SRT) to diversity and BOD removal, we constructed a composite variable of ‘SRT effect’ as a linear combination of
SRT and the square of SRT (SRT.SQ). Observed variables are represented by rectangles, and the composite variable is represented by a hexagon. The community composition is represented by the first principal component score (PC1) from the Bray-Curtis distance-based principal coordinate analysis. Blue and red arrows represent significant positive and negative pathways, respectively. Numbers near the pathway arrow indicate the standard path coefficients (β). Arrow width is proportional to the strength of the relationship. \( R^2 \) next to boxes represents the proportion of variance explained for every dependent variable. All the pathways in this model were significant (P < 0.05). Model \( \chi^2 = 13.92, df = 12, P = 0.31, N = 1186; \) root mean square error of approximation (RMSEA) = 0.012 with probability of a close fit = 1.00. (c) The average rRNA gene copy number of the community increased with the influent BOD/(1+recycle ratio) which approximates the influent BOD level of aerobic tank. The rRNA gene copy number is a functional trait reflecting the life-history strategy of microorganisms. (d) The strength of association between taxonomic composition and temperature was tested by random forest. The prediction accuracy (represented by cross-validated \( R^2 \)) is used as a measure for the strength of association. The red diagonal shows the theoretical curve for perfect predictions. As an example, we show (inset) that models trained on data from North and South America samples could achieve a moderate prediction accuracy on temperature in Asian samples.
Supplementary Materials for

Global Diversity and Biogeography of the Bacterial Communities That Clean Our Wastewater

Linwei Wu, Daliang Ning, Bing Zhang, Yong Li, Ping Zhang, Xiaoyu Shan, Qiuting Zhang, Mathew Brown, Zhenxin Li, Joy D. Van Nostrand, Fangqiong Ling, Naijia Xiao, Julia Vierheilig, George F. Wells, Yunfeng Yang, Ye Deng, Qichao Tu, Aijie Wang, Global Water Microbiome Consortium, Tong Zhang, Zhili He, Jurg Keller, Per H. Nielsen, Pedro J. J. Alvarez, Craig S. Criddle, Michael Wagner, James M. Tiedje, Qiang He, Thomas P. Curtis, David A. Stahl, Lisa Alvarez-Cohen, Bruce E. Rittmann, Xianghua Wen, and Jizhong Zhou

Correspondence to: jzhou@ou.edu; xhwen@tsinghua.edu.cn; tom.curtis@newcastle.ac.uk; qianghe@utk.edu.

This PDF file includes:

- Supplementary Figs. 1 to 7
- Supplementary Tables 1 to 5
Supplementary Figure 1 | Sequencing efforts and alpha diversity patterns. a, Summary of sampling and sequencing statistics. Sequencing effort was estimated before resampling, while the alpha diversity indices were calculated based on the resampled OTU table. G, giga, $10^9$; M, million, $10^6$; K, thousand, $10^3$; reads, detected sequence number; PD, Faith’s phylogenetic diversity$^{91}$. Values in some columns are presented as mean ± SD. b-e, Rarefaction curves of all sequences at sample, plant, continent and global levels. In total, 48,573,843 sequence reads were obtained after quality filtering, which are mapped to a total of 96,148 OTUs using UPARSE. After removing global singletons$^{59}$, the rarefaction curve at the global level appeared to almost
reach plateau and the curves for North America and Asia approached asymptotes. \( f \), Plotting phylogenetic diversity against the absolute latitude of sampling locations showed the peak of diversity at the intermediate latitude (\( n = 1186 \)). The color gradient denotes the annual mean air temperature. Shapes of symbols denote whether a sample originated from Northern (circle) or Southern Hemisphere (square). \( g-h \), Relationships between bacterial richness and mean annual air temperature (\( n = 1,186 \); \( g \)) or mixed liquid temperature (\( n = 787 \); \( h \)). The natural log of richness was used for analyzing the relationships between OTU richness and temperature, which was expressed as the inverse of temperature in kelvin. In fact, there is strong correlation between air temperature and mixed liquid temperature (\( n = 787 \); insert of \( g \)). The small \( R^2 \) values indicate a low explanatory power of the metabolic theory of ecology (MTE)\(^{39} \).
Supplementary Figure 2 | Summary of metadata. The 1186 AS samples were taken from 269 WWTPs across six continents and represent diverse geographies, operation conditions and chemistries. The WWTPs include those built as early as 1913 to new systems built in the 2010s, with treatment capacity (influent flow rate) ranging from 450 to $3.6 \times 10^6$ m$^3$ per day, influent BOD levels from 30 to 550 mg/L and solids retention time (SRT) from 2 to 30 days for most WWTPs. The nutrient level of the influent municipal sewage varied. For example, the influent total nitrogen (TN) ranged from 0.33 to 148 mg N/L, and the influent total phosphorus (TP) ranged from 0.6 to 22 mg P/L. Detailed metadata for each WWTP are shown in Supplementary Table 1.
Supplementary Figure 3 | Species abundance distributions and beta diversity patterns. 

a. Comparison of the performance of SAD models for AS bacterial communities. For each model, the mean and standard deviation of modified $r^2 (r^2_m)$ from the SAD model fitting on 1186 AS samples are shown. The log-normal provided the best predictions for how abundance varies among different taxa. 

b. Occupancy-frequency and occupancy-abundance distributions of all OTUs. OTUs were classified into bins with every 5% as interval according to their occupancy of WWTPs. For example, the first bin contains OTUs that were detected in less than 5% of the plants, while the last bin contains OTUs that were detected in 95%-100% of the plants. For each bin, species frequency, that is, the number of OTUs was counted and indicated by the blue bar in this figure; the total abundance of OTUs in each bin, was also calculated and is indicated by the red point. Only 0.14% of OTUs were found in more than 95% (i.e., more than 256 out of 269) of...
the WWTPs, which represented 13.61% of the total abundance. In contrast, the majority of microbial OTUs (71.90%) were only found in less than 5% (i.e., less than 13 out of 269) of the WWTPs, which only represented 4.77% of the total sequence reads. Similar hyper-dominance patterns have been observed for both macro- and microbial communities. For instance, in the Amazonia ecosystem, 227 tree species out of a total of 16,000 (1.42%) accounted for half of all trees. In Tara oceans, 0.25% of the eukaryotic marine plankton OTUs represented 48% of all rRNA gene reads. Our study shows that 866 microbial OTUs (1.41%) represented 50.06% of the total abundance in AS of WWTPs, a proportion very similar to what observed in the Amazonia tree community.

c. Phylum-level (class-level for Proteobacteria) community compositions across cities. Each column represents a city, which is clustered based on the Bray-Curtis similarity of phylum-level compositions. The country is labeled with text on the bottom, while the continent of the site is indicated with colored circle on the top.

d-e. Principal coordinate analysis (PCoA) showed variations in bacterial taxonomic composition (Bray-Curtis distance; d) and phylogenetic composition (weighted UniFrac; e) across continents. Lines above and to the right of PCoA plots show the density distribution of PC1 and PC2 scores for each continent, indicating distinctions between continents. For example, the density plots in e show distinct peaks in PC2 for Asia, Europe, and North America.

f. The upper triangle (grey shaded) shows the F values of PERMANOVA (Permutational multivariate analysis of variance using distance matrices) on taxonomic composition dissimilarities while the lower triangle shows those of phylogenetic dissimilarities between continents. Significance level: * P < 0.05, ** P < 0.01, *** P < 0.001.

g. PERMANOVA showing significant effects of continents, climate types, activated sludge process types and their interactions on bacterial community compositions. Df, the degree of freedom; SumSqs, the sum of squares due to the source; MeanSqs, the mean sum of squares due to the source; F.Model, the F-statistic. The effect of continent overwhelmed the effects of climate types or AS process types, as indicated by the highest mean square values in the PERMANOVA models.

h. Comparing bacterial taxonomic compositions (Bray-Curtis distance) across climate types.

i. Comparing bacterial phylogenetic compositions (weighted UniFrac) across climate types. The detailed classification of climate types is shown in Supplementary Fig. 7.
Supplementary Figure 4 | Scale-dependent distance-decay patterns. a, Distance-decay relationships (DDRs) of taxonomic and phylogenetic similarities. For each similarity index, the DDR was significant at the overall scale, within-city and across city (within a continent) scale, but not at the across continent scale. b, Relative importance of geographic distance and environmental variables by Variation Partition Analysis (VPA). The results based on Sørensen dissimilarity, weighted UniFrac, and unweighted UniFrac are presented here. The result based
on Bray-Curtis are presented in Fig. 3c and 3d. Our results suggest that, as spatial scales increase, geographic effects decrease, but environmental effects increase. At the across-continental scale, the contribution of geographical distance in explaining bacterial community dissimilarity is very little (<0.1%). One explanation is a dispersal barrier at large scale, e.g., between continents. Without dispersal, drift would create a patchy distribution of microorganisms. Thus, when considering samples from different continents, an increase in geographical distance does not contribute to higher microbial β-diversity. The distance effect, which quantifies the correlation between geographical distance and microbial β-diversity, becomes smaller or indeed there is no effect at larger scales. Alternatively, but not exclusively, unmeasured environmental variables may be spatially auto-correlated at small scales. For example, if the unmeasured variable salinity is spatially auto-correlated within a city, but not across cities, we could overestimate the distance effects at the within-city scale, which should be attributed to the environmental effects. It should be noted that the sequencing depth and the taxonomic resolution of the phylogenetic markers used also could explain community variations observed to some degree.
**Supplementary Figure 5 | Correlation between bacterial community composition and AS functions.**

**a.** Partial Mantel tests showing significant correlations between the distance-corrected dissimilarities of bacterial community composition and those of AS functions representing by chemical removal rates. MLSS, mixed liquor suspended solids, related to microbial biomass. Enhanced biological phosphorus removal (EBPR) plants included those with the anaerobic tank(s). We also tested the correlation between ammonia-nitrogen removal rate and the nitrification community which only included Nitrospira and Nitrosomonas OTUs.

**b.** Correlation between the abundance of typical phylotypes and AS functions. The color gradient on the right indicates Spearman’s correlation coefficients, and the asterisks denote the significance levels: *** $P < 0.001$, ** $P < 0.01$, * $P < 0.05$. In the correlation analyses, $P$-values were adjusted for multiple testing using the Benjamini and Hochberg false discovery rate (FDR) controlling procedure, where all phyla (Proteobacteria at class level) or genera were included. Only phyla presented in at least 60% of samples, as well the genera Nitrospira and Nitrosomonas are shown. Both Nitrospira and Nitrosomonas were positively correlated with ammonia-nitrogen removal percentage but not ammonia-nitrogen removal rate, which might imply a trade-off between ‘good’ (removal percentage) and ‘fast’ (removal rate) AS performance. In fact, Nitrospira is the top genus correlating to ammonia-nitrogen removal percentage.

**c.** Histogram showing the distribution of Spearman’s correlation coefficients between non-core OTUs and AS functions (chemical removal rates). In the correlation analysis, all OTUs detected in at least 20% of samples were included, and $P$-values were adjusted by FDR controlling procedure. Only significant correlations ($P < 0.05$) were included. Correlations for core OTUs are shown in Figure 2b.
Supplementary Figure 6 | Strength of association between bacterial and environmental data tested by random forest.  

a. To control for the strong autocorrelation among samples from the same plant, random forest models were constructed at the plant level, using average species abundance data among samples within the same plant. Our baseline model used the full data-set for training. We subsequently trained new random forests for each sample with customized training sets that excluded samples within a defined radius of the target sample. We varied the size of this radius of exclusion from 0 to 5000 km. The prediction strength (represented by cross-validated $R^2$) decreases as the radius increases, but mean air temperature (MAT) is the most robust. 
b. Heatmap of the top 50 OTUs which were important in predicting MAT in the baseline random forest model. In the heatmap, each column represents a plant, which is ranged by MAT. The blue color gradients indicate the log abundance of OTUs. There’re clear patterns of OTU dynamics in response to temperature. For example, three δ-proteobacteria OTUs (OTU_212, OTU_621 and OTU_416) were more abundant under intermediate to high temperatures than low temperatures.
Supplementary Figure 7 | World map of Köppen-Geiger climate classification\(^53\). Color indicates climate types: Af, Tropical rainforest climate; Am, Tropical monsoon climate; Aw, Tropical savanna climate; BWh, Hot desert climate; BWk, Cold desert climate; BSh, Hot semi-arid climate; BSk, Cold semi-arid climate; Csa, Hot summer Mediterranean climate; Csb, Warm-summer Mediterranean climate; Cwa, Humid subtropical climate; Cwb, Subtropical highland climate; Cwc, Cold subtropical highland climate; Cfa, Humid subtropical climate; Cfb, Temperate oceanic climate; Cfc, Subpolar oceanic climate; Dsa, Mediterranean-influenced hot-summer humid continental climate; Dsb, Mediterranean-influenced warm-summer humid continental climate; Dsc, Mediterranean-influenced subarctic climate; Dsd, Mediterranean-influenced extremely cold, subarctic climate; Dwa, Monsoon-influenced hot-summer humid continental climate; Dwbl, Monsoon-influenced warm-summer humid continental climate; Dwc, Monsoon-influenced subarctic climate; Dwd, Monsoon-influenced extremely cold subarctic climate; Dfa, Hot-summer humid continental climate; Dfb, Warm-summer humid continental climate; Dfc, Subarctic climate; Dfd, Extremely cold subarctic climate; ET, Tundra climate; and EF, Ice cap climate.
Supplementary Tables

Supplementary Table 1. Summary of metadata.
See the excel file.

Supplementary Table 2. List of core OTUs.
See the excel file.
Supplementary Table 3. OTUs identified as members of the core community for each habitat based on the merged OTU table of our AS samples and samples from EMP. The first column shows the OTU id of Greengenes database and the second column shows the lowest taxonomy level the OTU could be classified to, where o, f, g, and s represent order, family, genus, and species, respectively. The third to the last column show the average relative abundance and occupancy of samples for each habitat, in which data are presented as relative abundance (occupancy); Core OTUs are indicated by color for each column. For example, core OTUs for AS in WWTPs are those filled with pink.

<table>
<thead>
<tr>
<th>Id</th>
<th>taxonomy</th>
<th>WWTP</th>
<th>Air</th>
<th>Animal feces</th>
<th>Freshwater</th>
<th>Human feces</th>
<th>Ocean</th>
<th>Soil</th>
</tr>
</thead>
<tbody>
<tr>
<td>4468234</td>
<td>g: <em>Bacteroides</em></td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>8.06%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4467447</td>
<td>g: <em>Bacteroides</em></td>
<td>0.00%</td>
<td>0.61%</td>
<td>0.81%</td>
<td>4.02%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4453524</td>
<td>f: <em>Comamonadaceae</em></td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.86%</td>
<td>0.00%</td>
<td>0.09%</td>
<td>0.15%</td>
</tr>
<tr>
<td>4457032</td>
<td>g: <em>DA101</em></td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>0.05%</td>
<td>0.94%</td>
</tr>
<tr>
<td>4377104</td>
<td>f: <em>Bradyrhizobiaceae</em></td>
<td>0.00%</td>
<td>0.16%</td>
<td>0.01%</td>
<td>0.14%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.79%</td>
</tr>
<tr>
<td>4422313</td>
<td>o: <em>Streptophyta</em></td>
<td>0.00%</td>
<td>2.52%</td>
<td>0.52%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.02%</td>
<td>0.02%</td>
</tr>
<tr>
<td>4447072</td>
<td>g: <em>Bacteroides</em></td>
<td>0.00%</td>
<td>0.45%</td>
<td>0.25%</td>
<td>0.00%</td>
<td>2.70%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>228207</td>
<td>f: <em>Moraxellaceae</em></td>
<td>1.57%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>819699</td>
<td>o: <em>Saprospirales</em></td>
<td>1.49%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>548280</td>
<td>f: <em>Xanthomonadaceae</em></td>
<td>1.49%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>570443</td>
<td>o: <em>Mycococcales</em></td>
<td>1.43%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4481131</td>
<td>s: <em>Faecalibacterium prausnitzii</em></td>
<td>0.00%</td>
<td>1.26%</td>
<td>0.08%</td>
<td>0.00%</td>
<td>2.30%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>584010</td>
<td>g: <em>Zoogloeae</em></td>
<td>1.36%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1108726</td>
<td>g: <em>Methylibium</em></td>
<td>1.30%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4342972</td>
<td>g: <em>DA101</em></td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.19%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>719367</td>
<td>f: <em>Comamonadaceae</em></td>
<td>1.08%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>879612</td>
<td>f: <em>Rhodocyclaceae</em></td>
<td>1.18%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>551895</td>
<td>f: <em>Chitinophagaceae</em></td>
<td>1.15%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4465907</td>
<td>f: <em>Lachnospiraceae</em></td>
<td>0.99%</td>
<td>0.37%</td>
<td>0.09%</td>
<td>0.00%</td>
<td>2.24%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4308759</td>
<td>f: <em>Syntrophobacteraceae</em></td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.00%</td>
<td>0.12%</td>
<td>0.00%</td>
<td>0.01%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1108275</td>
<td>g: <em>Limnophitans</em></td>
<td>1.03%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4468097</td>
<td>o: <em>iii1-15</em></td>
<td>0.00%</td>
<td>0.04%</td>
<td>0.08%</td>
<td>0.08%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4460870</td>
<td>g: <em>Nitrospirae</em></td>
<td>0.98%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>4479989</td>
<td>g: <em>Streptococcus</em></td>
<td>0.00%</td>
<td>2.71%</td>
<td>0.03%</td>
<td>0.01%</td>
<td>0.04%</td>
<td>0.02%</td>
<td>0.02%</td>
</tr>
<tr>
<td>4457438</td>
<td>f: <em>Lachnospiraceae</em></td>
<td>0.00%</td>
<td>0.61%</td>
<td>0.02%</td>
<td>0.00%</td>
<td>1.79%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>1105514</td>
<td>g: <em>Dechloromonas</em></td>
<td>0.87%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
<td>0.00%</td>
</tr>
<tr>
<td>Accession</td>
<td>Family</td>
<td>Genus</td>
<td>Percentage (Identity)</td>
<td>Percentage (Identity)</td>
<td>Percentage (Identity)</td>
<td>Percentage (Identity)</td>
<td>Percentage (Identity)</td>
<td>Percentage (Identity)</td>
</tr>
<tr>
<td>-----------</td>
<td>--------</td>
<td>-------</td>
<td>-----------------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
<td>-----------------------</td>
</tr>
<tr>
<td>1104733</td>
<td>f: Comamonadaceae</td>
<td><strong>0.83%</strong> (100.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>4373617</td>
<td>f: Sinobacteraceae</td>
<td>0.00% (0.00%)</td>
<td>0.01% (12.35%)</td>
<td>0.00% (1.93%)</td>
<td>0.02% (29.04%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (4.95%)</td>
<td>0.28% (89.14%)</td>
</tr>
<tr>
<td>888396</td>
<td>g: Hydrogenophaga</td>
<td><strong>0.66%</strong> (100.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>4417475</td>
<td>o: Ellin329</td>
<td>0.00% (0.37%)</td>
<td>0.01% (7.41%)</td>
<td>0.00% (0.96%)</td>
<td>0.10% (39.37%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (5.37%)</td>
<td>0.13% (80.43%)</td>
</tr>
<tr>
<td>4373456</td>
<td>o: RB41</td>
<td>0.00% (0.00%)</td>
<td>0.02% (11.11%)</td>
<td>0.00% (0.32%)</td>
<td>0.04% (25.27%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (4.02%)</td>
<td>0.19% (80.43%)</td>
</tr>
<tr>
<td>813690</td>
<td>f: Comamonadaceae</td>
<td>0.01% (31.00%)</td>
<td>0.01% (13.58%)</td>
<td>0.00% (0.80%)</td>
<td>0.07% (48.48%)</td>
<td>0.00% (1.01%)</td>
<td>0.00% (6.71%)</td>
<td>0.15% (92.86%)</td>
</tr>
<tr>
<td>1106617</td>
<td>f: Comamonadaceae</td>
<td>0.56% (98.89%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>555423</td>
<td>g: Arcobacter</td>
<td>0.55% (96.68%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>993270</td>
<td>g: Dechloromonas</td>
<td>0.52% (94.83%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>831179</td>
<td>f: Comamonadaceae</td>
<td>0.47% (89.67%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.01% (24.42%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (2.06%)</td>
<td>0.00% (15.43%)</td>
</tr>
<tr>
<td>988314</td>
<td>s: Acinetobacter johnsonii</td>
<td>0.49% (96.31%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>745157</td>
<td>f: Rhodocyclaceae</td>
<td>0.49% (93.36%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.83%)</td>
<td>0.00% (0.57%)</td>
</tr>
<tr>
<td>4017244</td>
<td>g: Turneriella</td>
<td>0.43% (94.10%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.01% (23.82%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (1.96%)</td>
<td>0.01% (48.43%)</td>
</tr>
<tr>
<td>1087471</td>
<td>o: Sphingobacteriales</td>
<td>0.36% (88.56%)</td>
<td>0.00% (2.47%)</td>
<td>0.00% (0.80%)</td>
<td>0.02% (35.24%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (2.79%)</td>
<td>0.01% (24.14%)</td>
</tr>
<tr>
<td>4396984</td>
<td>f: Neisseriaceae</td>
<td>0.45% (89.30%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.16%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>1052264</td>
<td>g: Cloacibacterium</td>
<td>0.43% (99.26%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>942852</td>
<td>g: Aquabacterium</td>
<td>0.42% (97.42%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.16%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.61%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>807187</td>
<td>f: Saprospiraceae</td>
<td>0.41% (82.66%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>940737</td>
<td>s: Aquincola tertiaricarbonis</td>
<td>0.40% (100.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
<tr>
<td>810167</td>
<td>g: Methylibium</td>
<td>0.38% (96.31%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.24%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.14%)</td>
</tr>
<tr>
<td>1132602</td>
<td>f: Xanthomonadaceae</td>
<td>0.37% (86.72%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.85%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.10%)</td>
<td>0.00% (0.14%)</td>
</tr>
<tr>
<td>4441578</td>
<td>f: Halomonadaceae</td>
<td>0.36% (84.87%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.85%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
<td>0.00% (0.00%)</td>
</tr>
</tbody>
</table>
Supplementary Table 4. Statistical differences of community structure among different habitats. The upper triangle (grey shaded) shows the F values of PERMANOVA on Bray-Curtis dissimilarities, and the lower triangle shows the F values of PERMANOVA on Sorensen dissimilarities. The asterisks indicate the significance level: * P<0.05, ** P<0.01, *** P<0.001.

<table>
<thead>
<tr>
<th></th>
<th>WWTP</th>
<th>Air</th>
<th>Animal feces</th>
<th>Freshwater</th>
<th>Human feces</th>
<th>Ocean</th>
<th>Soil</th>
</tr>
</thead>
<tbody>
<tr>
<td>WWTP</td>
<td>102.0***</td>
<td>174.2***</td>
<td>172.7***</td>
<td>137.9***</td>
<td>198.5***</td>
<td>190.5***</td>
<td></td>
</tr>
<tr>
<td>Air</td>
<td>109.0***</td>
<td></td>
<td>24.3***</td>
<td>37.5***</td>
<td>28.0***</td>
<td>38.8***</td>
<td>42.3***</td>
</tr>
<tr>
<td>Animal feces</td>
<td>192.2***</td>
<td>23.8***</td>
<td></td>
<td>101.5***</td>
<td>36.8***</td>
<td>119.8***</td>
<td>104.6***</td>
</tr>
<tr>
<td>Freshwater</td>
<td>186.3***</td>
<td>38.6***</td>
<td>117.3***</td>
<td></td>
<td>63.9***</td>
<td>115.2***</td>
<td>77.2***</td>
</tr>
<tr>
<td>Human feces</td>
<td>153.4***</td>
<td>28.9***</td>
<td>38.0***</td>
<td>71.8***</td>
<td></td>
<td>60.6***</td>
<td>72.7***</td>
</tr>
<tr>
<td>Ocean</td>
<td>220.0***</td>
<td>40.7***</td>
<td>142.5***</td>
<td>132.9***</td>
<td>67.7***</td>
<td></td>
<td>112.9***</td>
</tr>
<tr>
<td>Soil</td>
<td>214.7***</td>
<td>46.5***</td>
<td>125.0***</td>
<td>85.0***</td>
<td>85.0***</td>
<td>145.9***</td>
<td></td>
</tr>
</tbody>
</table>
Supplementary Table 7. Partial Mantel tests showing the correlations of the distance-corrected dissimilarities of bacterial community composition and those of environmental variables.

<table>
<thead>
<tr>
<th></th>
<th>Bray-Curtis distance</th>
<th>Sorensen distance</th>
<th>Weighted UniFrac</th>
<th>Unweighted UniFrac</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>r_m</td>
<td>P value</td>
<td>r_m</td>
<td>P value</td>
</tr>
<tr>
<td>Absolute latitude (°) (n=1186)</td>
<td>0.299</td>
<td>0.001</td>
<td>0.301</td>
<td>0.001</td>
</tr>
<tr>
<td>Mean annual air temperature (°C) (n=1186)</td>
<td>0.281</td>
<td>0.001</td>
<td>0.271</td>
<td>0.001</td>
</tr>
<tr>
<td>Influent COD (mg/L) (n=609)</td>
<td>0.277</td>
<td>0.001</td>
<td>0.313</td>
<td>0.001</td>
</tr>
<tr>
<td>SRT (days) (n=844)</td>
<td>0.268</td>
<td>0.001</td>
<td>0.273</td>
<td>0.001</td>
</tr>
<tr>
<td>Influent BOD (mg/L) (n=880)</td>
<td>0.225</td>
<td>0.001</td>
<td>0.245</td>
<td>0.001</td>
</tr>
<tr>
<td>Influent TP (mg/L) (n=603)</td>
<td>0.031</td>
<td>0.063</td>
<td>0.037</td>
<td>0.031</td>
</tr>
<tr>
<td>Influent TN (mg/L) (n=479)</td>
<td>0.002</td>
<td>0.415</td>
<td>0.006</td>
<td>0.367</td>
</tr>
<tr>
<td>Influent NH₄-N (mg/L) (n=756)</td>
<td>-0.040</td>
<td>1</td>
<td>-0.043</td>
<td>1</td>
</tr>
<tr>
<td>Effluent BOD (mg/L) (n=848)</td>
<td>0.176</td>
<td>0.001</td>
<td>0.184</td>
<td>0.001</td>
</tr>
<tr>
<td>Effluent COD (mg/L) (n=565)</td>
<td>0.330</td>
<td>0.001</td>
<td>0.371</td>
<td>0.001</td>
</tr>
<tr>
<td>Effluent TP (mg/L) (n=611)</td>
<td>0.117</td>
<td>0.001</td>
<td>0.131</td>
<td>0.001</td>
</tr>
<tr>
<td>Effluent TN (mg/L) (n=469)</td>
<td>0.113</td>
<td>0.001</td>
<td>0.117</td>
<td>0.001</td>
</tr>
<tr>
<td>Effluent NH₄-N (mg/L) (n=757)</td>
<td>0.183</td>
<td>0.001</td>
<td>0.180</td>
<td>0.001</td>
</tr>
<tr>
<td>Min annual air temperature (°C) (n=1186)</td>
<td>0.200</td>
<td>0.001</td>
<td>0.203</td>
<td>0.001</td>
</tr>
<tr>
<td>pH (n=869)</td>
<td>0.179</td>
<td>0.001</td>
<td>0.176</td>
<td>0.001</td>
</tr>
<tr>
<td>F/M ratio (kg BOD/kg MLSS-d) (n=541)</td>
<td>0.162</td>
<td>0.001</td>
<td>0.159</td>
<td>0.001</td>
</tr>
<tr>
<td>Mixed liquid temperature (n=787)</td>
<td>0.153</td>
<td>0.001</td>
<td>0.143</td>
<td>0.001</td>
</tr>
<tr>
<td>DO (mg/L) (n=927)</td>
<td>0.130</td>
<td>0.001</td>
<td>0.131</td>
<td>0.001</td>
</tr>
<tr>
<td>Industrial source percentage (n=707)</td>
<td>0.123</td>
<td>0.001</td>
<td>0.129</td>
<td>0.001</td>
</tr>
<tr>
<td>Air temperature of sampling month (n=1042)</td>
<td>0.113</td>
<td>0.001</td>
<td>0.116</td>
<td>0.001</td>
</tr>
<tr>
<td>BOD/(1+recycle ratio) (mg/L) (n=641)</td>
<td>0.111</td>
<td>0.001</td>
<td>0.111</td>
<td>0.001</td>
</tr>
<tr>
<td>Annual precipitation (mm) (n=1186)</td>
<td>0.081</td>
<td>0.001</td>
<td>0.083</td>
<td>0.001</td>
</tr>
<tr>
<td>Precipitation of sampling month (n=1163)</td>
<td>0.080</td>
<td>0.001</td>
<td>0.087</td>
<td>0.001</td>
</tr>
<tr>
<td>Variable</td>
<td>Coefficients</td>
<td>p-values</td>
<td>Coefficients</td>
<td>p-values</td>
</tr>
<tr>
<td>------------------------------------------------------------</td>
<td>--------------</td>
<td>----------</td>
<td>--------------</td>
<td>----------</td>
</tr>
<tr>
<td>GDP (per capita dollars) (n=1140)</td>
<td>0.062</td>
<td>0.001</td>
<td>0.060</td>
<td>0.001</td>
</tr>
<tr>
<td>MLSS (mg/L) (n=765)</td>
<td>0.052</td>
<td>0.012</td>
<td>0.051</td>
<td>0.023</td>
</tr>
<tr>
<td>Max annual air temperature (°C) (n=1186)</td>
<td>0.052</td>
<td>0.001</td>
<td>0.052</td>
<td>0.005</td>
</tr>
<tr>
<td>Plant age (years) (n=930)</td>
<td>0.050</td>
<td>0.002</td>
<td>0.046</td>
<td>0.111</td>
</tr>
<tr>
<td>Tank HRT (h) (n=634)</td>
<td>0.034</td>
<td>0.049</td>
<td>0.034</td>
<td>0.054</td>
</tr>
<tr>
<td>COD/(1+recycling ratio) (mg/L) (n=437)</td>
<td>0.018</td>
<td>0.166</td>
<td>0.021</td>
<td>0.148</td>
</tr>
<tr>
<td>Influent BOD/COD ratio (n=522)</td>
<td>0.012</td>
<td>0.268</td>
<td>0.002</td>
<td>0.463</td>
</tr>
<tr>
<td>Designed capacity (m³/d) (n=972)</td>
<td>-0.011</td>
<td>1</td>
<td>-0.009</td>
<td>1</td>
</tr>
<tr>
<td>Actual influent flow rate (m³/d) (n=967)</td>
<td>-0.020</td>
<td>1</td>
<td>-0.018</td>
<td>1</td>
</tr>
<tr>
<td>Volume (m³) (n=850)</td>
<td>-0.031</td>
<td>1</td>
<td>-0.030</td>
<td>1</td>
</tr>
</tbody>
</table>