environmental microbiology

Environmental Microbiology (2019) 21(1), 182–196 doi:10.1111/1462-2920.14439

Intensive allochthonous inputs along the Ganges River and their effect on microbial community composition and dynamics

Si-Yu Zhang,[1] Despina Tsementzi,[1] Janet K. Hatt,[1] Aaron Bivins,[1] Nikunj Khelurkar,[1] Joe Brown,[1] Sachchida Nand Tripathi[2,3] and Konstantinos T. Konstantinidis[1,4]*

[1]*School of Civil and Environmental Engineering, Georgia Institute of Technology, Ford Environmental Science & Technology Building, Atlanta, GA, 30332, USA.*
[2]*Department of Civil Engineering, Indian Institute of Technology, Kanpur, UP, 208016, India.*
[3]*Center for Environmental Science and Engineering, Indian Institute of Technology, Kanpur, UP, 208016, India.*
[4]*School of Biological Sciences, Georgia Institute of Technology, Ford Environmental Sciences & Technology Building, Atlanta, Georgia, 30332, USA.*

Summary

Little is known about microbial communities in the Ganges River, India and how they respond to intensive anthropogenic inputs. Here we applied shotgun metagenomics sequencing to study microbial community dynamics and function in planktonic samples collected along an approximately 700 km river transect, including urban cities and rural settings in upstream waters, before and after the monsoon rainy season. Our results showed that 11%–32% of the microbes represented terrestrial, sewage and human inputs (allochthonous). Sewage inputs significantly contributed to the higher abundance, by 13-fold of human gut microbiome (HG) associated sequences and 2-fold of antibiotic resistance genes (ARGs) in the Ganges relative to other riverine ecosystems in Europe, North and South America. Metagenome-assembled genome sequences (MAGs) representing allochthonous populations were detectable and tractable across the river after 1–2 days of (downstream) transport (> 200 km apart). Only approximately 8% of these MAGs were abundant in U.S. freshwater ecosystems, revealing distinct biodiversity for the Ganges. Microbial communities in the rainy season exhibited increased alpha-diversity and spatial heterogeneity and showed significantly weaker distance-decay patterns compared with the dry season. These results advance our understanding of the Ganges microbial communities and how they respond to anthropogenic pollution.

Introduction

The Ganges River is the third largest river in the world by total amount of water discharged. It arises in the western Himalayas in the Indian state of Uttarakhand and traverses through many major cities such as Haridwar, Kanpur and Allahabad, draining about one quarter of India. Because of the geographical, historical, sociocultural and economic reasons, it is the most important river for the Indian people (Buhtiani *et al*., 2016). During the past couple of decades, intensified anthropogenic effects on the river have been reported, caused by expanding human population, industrialization and intensive agricultural practices (Sood *et al*., 2008). Direct waste discharge into the river from anthropogenic sources such as faecal (Tyagi *et al*., 2013), agricultural, industrial and sewage wastes (Namrata, 2010) has been reported. Moreover, poorly regulated use of antibiotics in India has resulted in highly antibiotic resistant bacterial pathogens (Murki, 2009), and threatens water quality in the Ganges River.

Allochthonous inputs from human sources can increase the abundance of antibiotic resistance genes (ARGs; Marti *et al*., 2013) or mobile elements (Gillings *et al*., 2015), with potentially important public health ramifications. Driven by the rapid advances in sequencing technology, metagenomic approaches have been widely applied to the study of microbial communities in riverine and other ecosystems (Oh *et al*., 2011; Ghai, 2011; Abia *et al*., 2018). Studies of anthropogenic effects on microbial communities in freshwater systems have shown that the discharge of industrial waste can affect microbial diversity both at taxonomic and functional composition

Received 23 July, 2018; revised 2 October, 2018; accepted 4 October, 2018. *For correspondence. E-mail kostas@ce.gatech.edu; Tel. +404-639-4292; Fax +404-894-8266.

levels (Chao *et al*., 2013). For instance, anthropogenic inputs along the Kalamas River in Greece can dramatically affect bacterial diversity with fewer than 5% of the total 16S rRNA gene-based operational taxonomic units (OTUs) being shared between samples from four seasons and two consecutive years (Meziti *et al*., 2016).

The microbial communities in the Ganges River and their responses to anthropogenic inputs remain essentially uncharacterized by culture-independent, metagenomic techniques, despite the apparent importance of the river for millions of Indians who live along its banks (Sharma *et al*., 2014). Recent studies have mostly focused on the isolation of pathogens or faecal coliform levels in the Ganges River (Tyagi *et al*., 2013; Hamner *et al*., 2007). It is also not clear how the Ganges microbial communities compare to their counterparts in other riverine ecosystems across the world. Compared with the other studied riverine ecosystems, the Ganges River in India has much higher density of human population along its banks, and presumably heavier contamination from industrial and human waste. Therefore, it represents an ideal ecosystem to study how freshwater microbes respond to anthropogenic inputs and potentially discover novel diversity. In addition, the Ganges covers a distance of over 2700 km, making it suitable to study the transportation of microbial populations along the river, especially between upstream, mountainous headwaters relative to downstream, density populated areas.

Herein, we investigate the variation in microbial communities in the Ganges River in two consecutive years, before or after the monsoon rainy season, at five sites along the river, aiming to address the following questions: (1) Are microbes in the Ganges river associated with high allochthonous inputs? (2) If yes, from where do allochthonous inputs originate and are they traceable along the Ganges River? (3) What is the effect of the monsoon (rainy season) and anthropogenic inputs on microbial community composition in the river?

Results

Description of samples and metagenomes

A total of 18 samples were collected along the Ganges River basin in India from upstream, less populated Gangotri and Haridwar areas, and downstream, urban settings in Kanpur, Allahabad and Agra (Fig. 1 and Supporting Information Table S1). Samples collected in 2015 were after the rainy season (region-wide monthly rainfall: 267.9 mm), that is, wet season samples. While samples collected in 2016 were before the rainy season, referred to as dry season samples (region-wide monthly rainfall: 19.1 mm) [rainfall statistics for India are available at http://hydro.imd.gov.in/hydrometweb/]. The wet season samples included those collected from upstream waters in Haridwar (Har_wet1 and Har_wet2), downstream

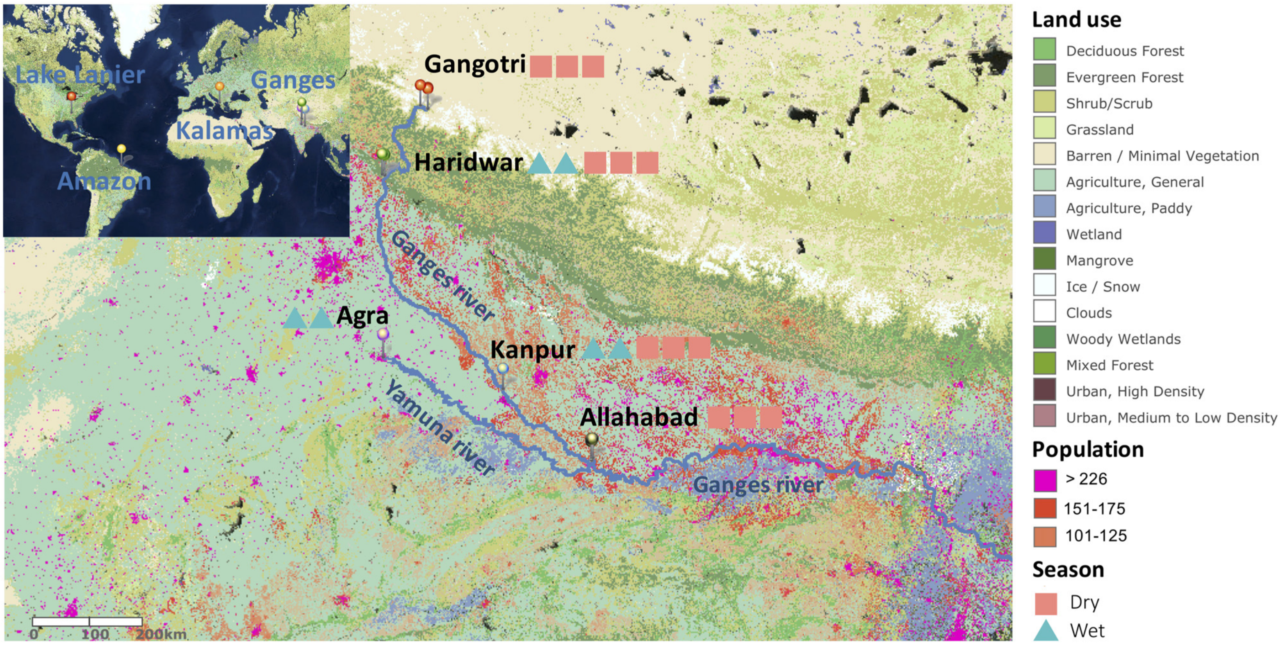

Fig. 1. Location of sampling sites along the Ganges River. Triangles represent samples collected in the wet season; squares represent samples collected in the dry season. The gradient from orange to pink indicates the population density along the river bank (see Figure key). The map of sampling sites was constructed on the ArcGIS platform and overlaid with World Land cover 30 m BaseVue 2013 and World Population Estimate [available at: http://www.arcgis.com/].

waters in Kanpur (Kan_wet1 and Kan_wet2), and a tributary to the Ganges River (Yamuna River) in Agra (Agr_wet1 and Agr_wet2). The dry season samples included those collected from upstream waters in Gangotri (Gan_dry1, Gan_dry2 and Gan_dry3) and Haridwar (Har_dry1, Har_dry2 and Har_dry3), and downstream waters in Kanpur (Kan_dry1, Kan_dry2 and Kan_dry3) and Allahabad (Alla_dry1, Alla_dry2 and Alla_dry3). Dry season samples showed substantial variation in their physiochemical properties. Notably, higher turbidity (1175 vs. 6–95) and lower pH (6.7 vs. 7.7–7.9) were detected upstream in Gangotri compared with the downstream (Supporting Information Table S2); no such measurements were performed in the wet season samples.

Around 5Gbp of shotgun metagenomic data were acquired after trimming for each of the Ganges samples. The coverage of the 18 metagenomes achieved by sequencing, that is, the fraction of the total extracted DNA that was sequenced, ranged from 45% to 85% as assessed by Nonpareil 3.0 (Rodriguez-R and Konstantinidis, 2014a, b; Rodriguez-R et al., 2018a, b). This level of coverage is adequate for between sample comparisons and assembly (Rodriguez-R and Konstantinidis, 2016). Between 183,096 and 431,188 genes were predicted from assembled contigs depending on the metagenome considered, with 30%–67% of total metagenomic reads mapping to the contigs (> 95% identity and > 50% query sequence coverage by the alignment; Supporting Information Table S3). 16S rRNA (16S) gene fragments extracted from the metagenomes were 0.1%–0.3% of the total reads as expected based on an average genome size of approximately 4 Mbp and 2–3 rRNA copies estimated for free-living bacteria. In each metagenome, 77%–90% of the 16S fragments were classified as bacterial at the class level, revealing that our metagenomic effort sampled predominantly bacteria.

To investigate the differences between the Ganges microbial communities and those from other freshwater ecosystems, one sample from the Amazon River in Brazil (Amazon), six samples from the Kalamas River in Greece (Kal1_Feb, Kal2_Feb, Kal1_May, Kal2_May, Kal1_Nov and Kal2_Nov), and three samples from the Chattahoochee River in the United States (Cha_Jan, Cha_Sep and Cha_Nov) were included in our comparative analysis. These samples were the only ones available from rivers with similar DNA sequence coverage and technology. The Amazon River sample was collected from a pristine area at a site nearly 400 km upstream from Manaus, Brazil (River delta), and the Chattahoochee River samples originated from Lake Lanier, which represents the upstream-most meromictic lake and the headwaters of the river. The Kalamas River samples originated from along the river and included waters that received treated and untreated sewage from the largest municipal city in the area.

Highly abundant allochthonous inputs in the Ganges River

To assess the allochthonous inputs in the Ganges River, 16S fragments were assigned to the different habitats (Fig. 2A) based on their best match analysis against a reference, in-house database (Meziti *et al*., 2016) [available at: http://enve-omics.ce.gatech.edu/data/]. About 82% ± 7% of the 16S-encoded reads were assignable to a habitat, with 25% ± 6% of them assigned to freshwater, 12% ± 6% assigned to other aquatic ecosystems, 11% ± 4% assigned to terrestrial, 6% ± 4% assigned to sludge/sewage, 0.6% ± 0.5% assigned to human related sequences (representing human gut, skin, oral, breast milk, etc.), and 27% ± 8% assigned as others, depending on the sample considered. Generally, less freshwater signal and higher abundance of allochthonous sequences/taxa from terrestrial and sludge/sewage were detected in the Ganges compared with the other freshwater ecosystems. For instance, the February samples from the Kalamas River had more than 40% of the total 16S sequences assigned to freshwater and less than 10% assigned to the terrestrial and sludge/sewage categories (Meziti *et al*., 2016).

Members of the *Proteobacteria* and *Actinobacteria* were the most abundant microbes in the Ganges, with 43% ± 21% and 15% ± 10% of the total 16S fragments classified to these phyla respectively. *Proteobacteria*, mostly comprised of *Betaproteobacteria* (average value of 40% ± 16% vs. 16% ± 3%) and *Gammaproteobacteria* (average value of 17% ± 17% vs. 3% ± 0.5%), were more dominant in samples from the upstream locations and wet season (Gan_dry1, Gan_dry2, Gan_dry3, Kan_wet1, Kan_wet2, Har_wet1 and Har_wet2) than in downstream samples from the dry season (Kan_dry1, Kan_dry2, Kan_dry3, Alla_dry1, Alla_dry2 and Alla_dry3). *Actinobacteria* were more abundant in downstream samples from the dry season than in upstream and wet season samples (27% ± 3% vs. 14% ± 6%) (Fig. 2B). Taxonomic assignment of protein-coding genes using MyTaxa (Luo *et al*., 2014) showed that allochthonous bacterial populations detected in the Ganges River in substantial *in-situ* relative abundance (typically > 1% of the total metagenomic reads; Supporting Information Fig. S1) included *Pseudomonas putida* (Nelson *et al*., 2002) and *Chthoniobacter flavus* (Sangwan *et al*., 2005), wastewater treatment plant-associated *Candidatus Nitrospira defluvii* (Maixner *et al*., 2008) and *Thauera* sp. MZ1T (Allen *et al*., 2004), pathogens *Acinetobacter baumannii* (Antunes et al., 2014), *Acinetobacter junii* (Bansal *et al*., 2017) and *Pseudomonas aeruginosa* (Stover *et al*., 2000), and antimicrobial-producing *Rheinheimera* sp. A13L (Gupta *et al*., 2011).

Ganges microbes harbour higher abundance of human gut signal and antibiotic resistance genes than counterparts elsewhere

Human gut associated microbes (HG) and antibiotic resistance gene (ARGs) sequence abundances in the Ganges River were further investigated, and compared with the other freshwater samples from the Amazon, Kalamas and Chattahoochee Rivers (Supporting Information Fig. S2) to assess the relative importance of anthropogenic inputs. Only the February samples from the Kalamas River, a medium to small sized river, were considered because it was in the middle of the rainy season and the water flow was comparable to those in the other, larger rivers. The average genome equivalents of HG sequences, that is, what fraction of the total cells sampled encoded the gene of interest, were about 13 times higher ($p < 0.05$), on average, in the Ganges samples collected in both the wet (all cells encoded 2.19 copies, on average) and dry seasons (1.98 copies per cell, on average) than the samples from the Amazon, Kalamas and Chattahoochee Rivers combined (0.17 copies per cell, on average) (Fig. 2C). The HG microbes in Ganges River were mostly classified as *Acinetobacter*, *Escherichia*, *Oxalobacer*, *Geobacter*, *Alistipes*, *Bacteroides*, *Prevotella* and *Caulobacter* at the genus level, accounting for 88% of the (classified) HG sequences (Supporting Information Fig. S3).

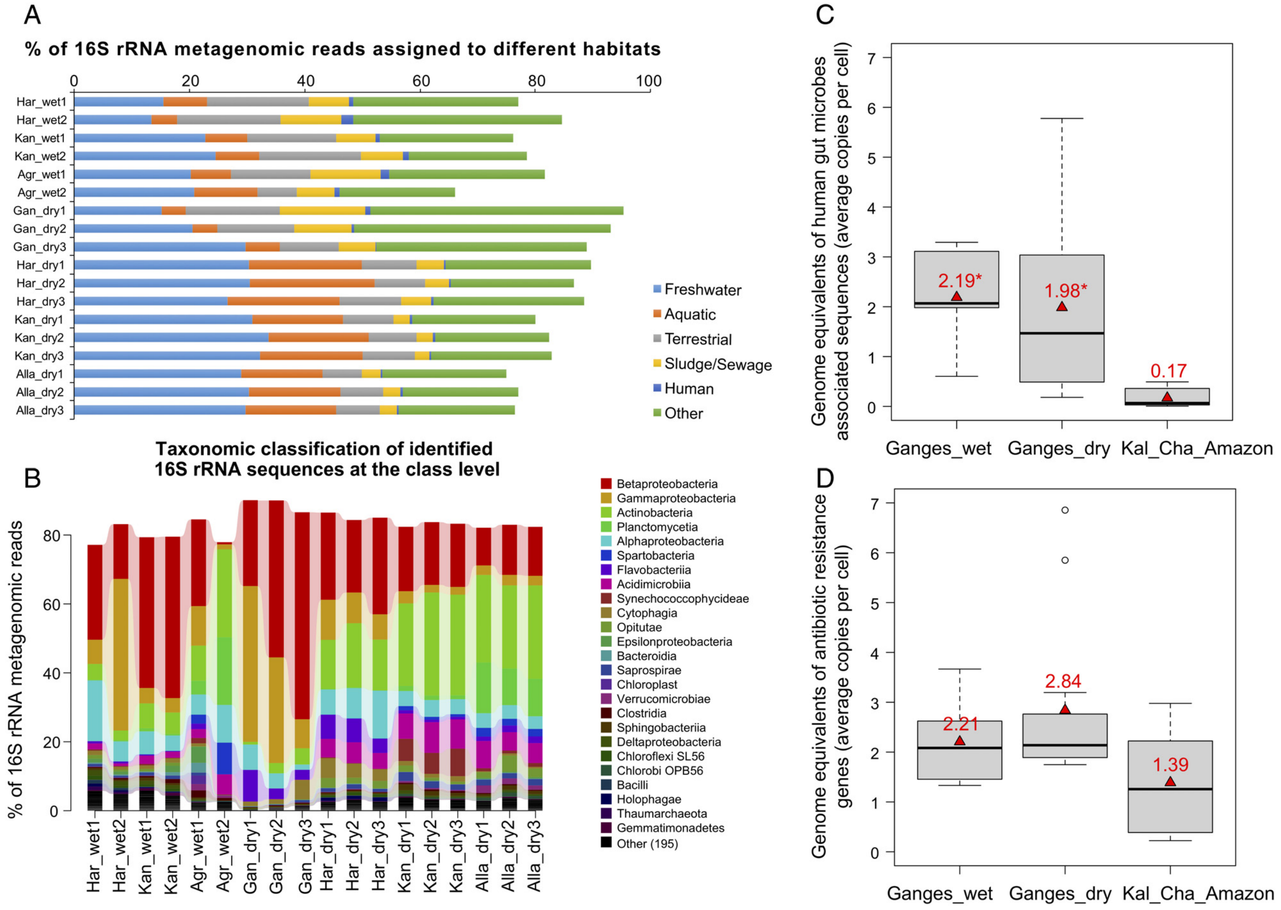

Fig. 2. Relative importance of allochthonous inputs based on habitat assignment of 16S rRNA and functional gene sequences.
A. Assignment of 16S rRNA gene sequence fragments to different habitats was based on best match analysis (> 99% nucleotide identity) against a reference *in-house* database (available at: http://enve-omisc.ce.gatech.edu/data/). The habitats including freshwater (lakes, rivers, streams), other aquatic (marine, estuaries, ground water, various aquatic environments), terrestrial (soil, plant and animal-associated sequences), sludge/sewage (sludge, sewage and wastewater-associated sequences), human (human related samples, representing human gut, skin, oral, breast milk, etc.) and others.
B. Class-level community composition and abundance of microbes in the Ganges River based on taxonomic classification of identified 16S rRNA gene sequence fragments. The relative abundance of each taxon at the class level was normalized by the total number of 16S rRNA gene sequence fragments obtained in each corresponding metagenome. Only the top 25 most abundant classes are shown.
C and D. Boxplots of genome equivalents (average copies per cell) of human gut microbiome associated sequences (C), and antibiotic resistance genes (D) in the Ganges River for both the wet (Ganges_wet) and dry (Ganges_dry) seasons, and in a combined category consisting of the other freshwater ecosystems including two samples from the Kalamas River, three samples from the Chattahoochee River, and one sample from the Amazon River (Kal_Cha_Amazon). Triangles display the mean value and horizontal lines display the median. * indicates p value < 0.05 as revealed by one-way ANOVA analysis (comparison of average genome equivalents of HG sequences in the Ganges River versus the other ecosystems).

The average genome equivalents of ARGs in the Ganges River (2.21 copies per cell for wet season and 2.84 copies per cell for dry season samples, on average) was about 2 times higher than those in the Amazon, Kalamas and Chattahoochee Rivers combined (1.39 copies per cell, on average) (Fig. 2D), albeit not statistically significant. Excluding samples from the Chattahoochee River, which had a relatively higher abundance of ARGs possibly resulting from anthropogenic inputs, rendered the abundance of ARGs to be significantly ($p < 0.05$) more abundant in Ganges River relative to other riverine ecosystems. Among the ARGs present in the Ganges, genes conferring resistance to aminoglycosides, trimethoprim, fluoroquinolones, polymyxin, chloramphenicol, tetracycline, phenicols, macrolides, aminocoumarins and beta-lactams were the most abundant, comprising about 93% of the ARGs annotated. These genes were detected in almost all the Ganges samples, but with a relatively higher abundance in the samples from upstream locations and/or the wet season (Supporting Information Fig. S4).

High abundance of HG signals and ARGs correlated with sludge/sewage inputs

Pearson correlation showed significant positive correlation of HG or ARGs relative abundance (genome equivalents) with the percentage of 16S gene-encoding reads assigned to sludge/sewage ($r = 0.74$, $p < 0.001$ and $r = 0.72$, $p < 0.001$, respectively). In contrast, significant negative correlations were identified for HG or ARGs relative abundance with the percentage of 16S-encoding reads assigned to freshwater origin ($r = 0.50$, $p < 0.05$ and $r = 0.53$, $p < 0.05$, respectively), and for ARGs with other aquatic ecosystems origin ($r = 0.53$, $p < 0.05$). No significant correlation was observed for HG or ARGs with 16S-encoding reads assigned to terrestrial or human, or for HG sequences with other aquatic ecosystems (Fig. 3). While only a weak correlation between HG and 16S-encoding reads assigned to human origin was observed when all samples were considered ($r = 0.36$, $p > 0.1$), when the analysis was restricted to samples with substantial HG signal (e.g., as opposed to signal below detection limit), a strong positive correlation was observed ($r = 0.77$, $p < 0.05$). Significant positive correlation was also observed between the relative abundance of HG and ARG sequences ($r = 0.68$, $p < 0.01$, Supporting Information Fig. S5).

Novelty of the Ganges microbial communities

Seventy-four percent of the 102 metagenome-assembled genomes or MAGs (Supporting Information Table S4)

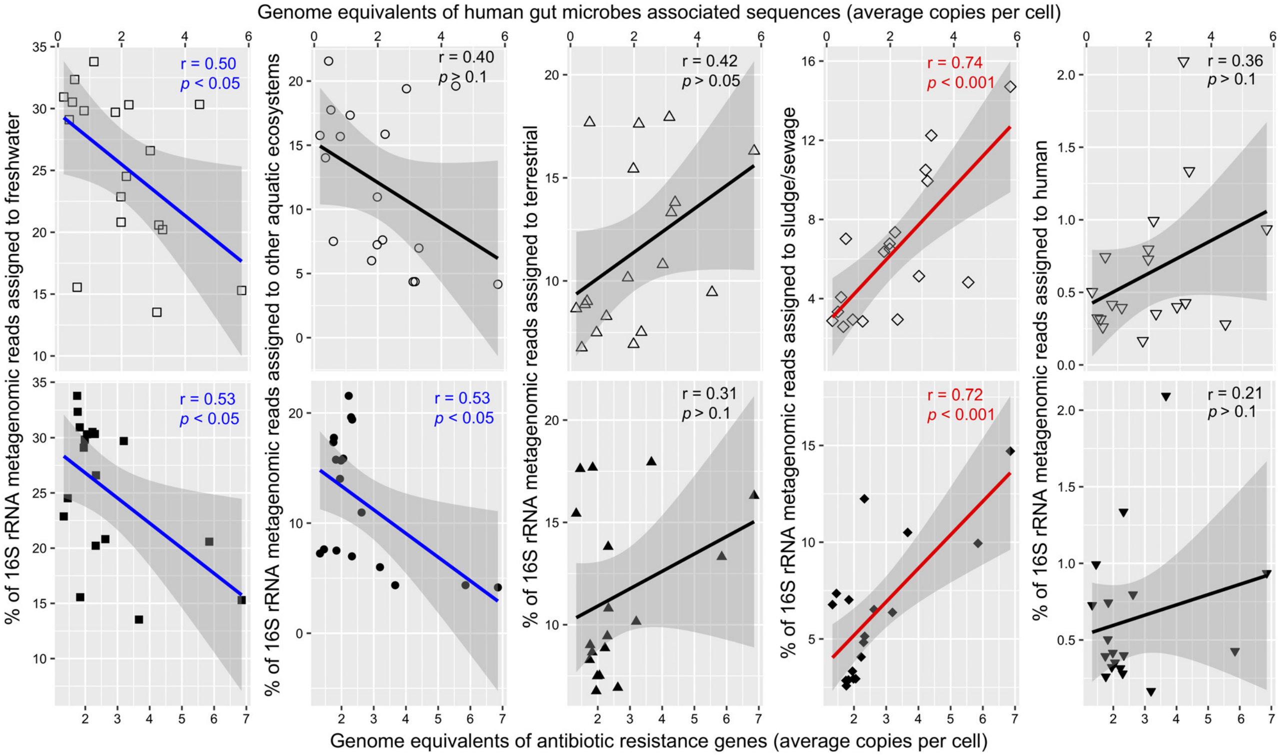

Fig. 3. Correlation of the genome equivalents of human gut microbes associated sequences and antibiotic resistance genes to the percentage of 16S rRNA gene sequence fragments assigned to different habitats. The fit-line in blue indicates significant negative correlation ($p < 0.05$); the fit-line in red indicates significant positive correlation ($p < 0.001$); the fit-line in black indicates no significant correlation was found ($p > 0.1$). [Color figure can be viewed at wileyonlinelibrary.com]

recovered from the 18 metagenomic datasets of the Ganges represented a novel species with probability $p \leq 0.01$ based on the MiGA classification (Rodriguez-R et al., 2018a, b). Fourteen percent of the remaining MAGs were novel at the genus level, 5% at family level and the last 8% represented members of previously described species. The majority of the 102 Ganges MAGs (n = 59, or around 60% of total) was assignable to *Actinobacteria* (n = 21), *Betaproteobacteria* (n = 16), *Chitinophagia* (n = 8), *Acidimicrobiia* (n = 5), *Alphaproteobacteria* (n = 2), *Verrucomicrobiae* (n = 2), *Sphingobacteriia* (n = 2), *Planctomycetia* (n = 2) and *Cytophagia* (n = 1) at the class level ($p \leq 0.01$). Members of *Comamonadaceae* (n = 4), *Bradyrhizobiaceae* (n = 2) and unclassified *Burkholderiales* (n = 2), etc were the most abundant families among the MAGs (n = 22, or around 20% of total) (Fig. 4A). Only around 8% of the MAGs (n = 8) were shared (i.e., > 95% genome-aggregate average nucleotide identity) with the 1126 MAGs recovered from Chattahoochee River. Among the eight MAGs, six of them originated from urban settings in Kanpur, and two of them were recovered from the less populated area in Haridwar. These MAGs most likely ($p \leq 0.01$) belong to *Betaproteobacteria* (n = 3), *Alphaproteobacteria* (n = 1), *Sphingobacteria* (n = 1) at the class level, and *Methylocystaceae* (n = 1), *Comamonadaceae* (n = 1) at the family level, and one of them was classified only at the domain level as Bacteria.

PCoA based on Mash distances of whole metagenomes (Fig. 4B) showed the separation pattern of Ganges microbial communities from their counterparts elsewhere confirming the MAG-based results reported above, with the possible exception of one sample from Gangotri (Gan_dry1) that clustered closely to one sample from Kalamas (Kal1_Nov). Interestingly, the latter Kalamas sample was previously reported to contain a relatively high abundance of sludge/sewage, and terrestrial associated sequences (Meziti *et al*., 2016), similar to what was observed for the Gangotri sample in this study.

Tracking individual allochthonous populations along the Ganges River

Two population genomes, that is, Gan_dry1.MAG001 and Gan_dry2.MAG001, which showed the highest relative abundance in, and were assembled from, the

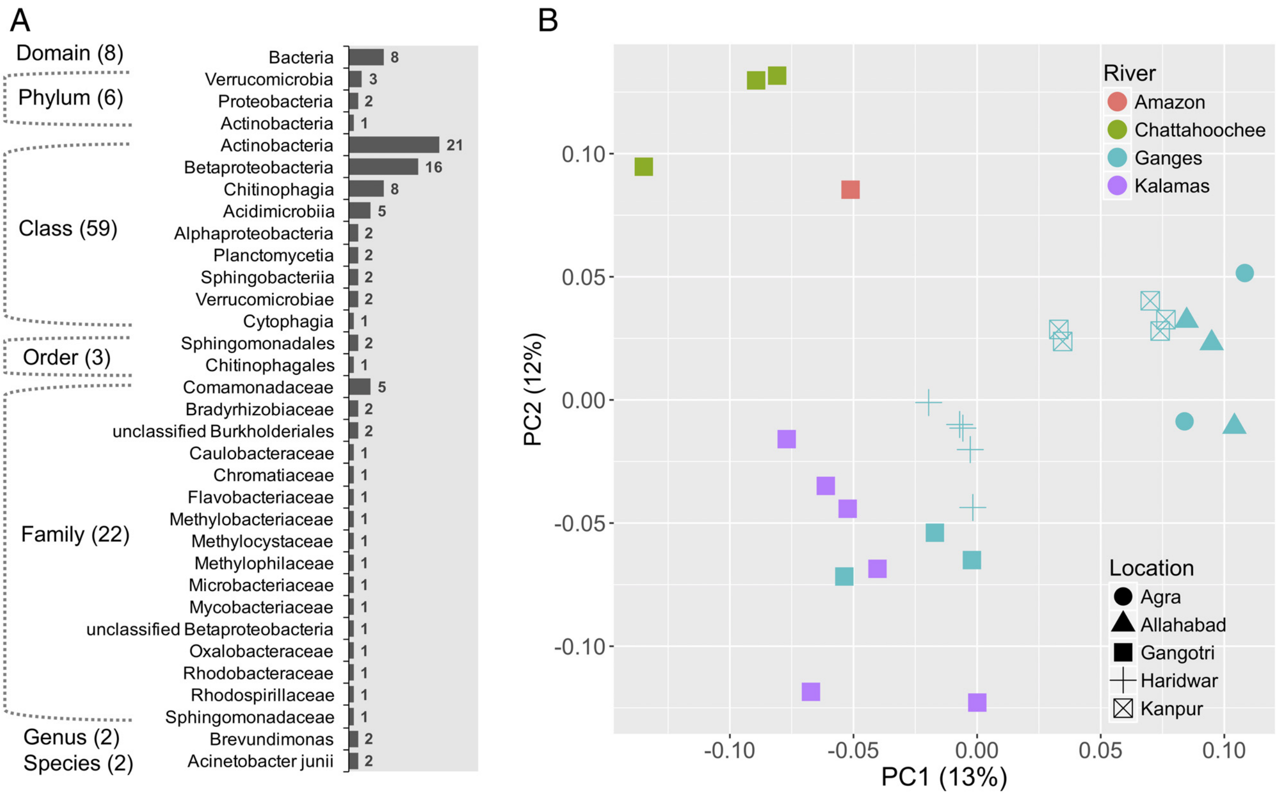

Fig. 4. Novelty of Ganges River microbial communities.
A. Summary of the taxonomy of 102 MAGs recovered from the 18 metagenomic datasets of Ganges based on MiGA. The graph shows the lowest possible classification ($p < 0.01$) that was achievable by MiGA for each MAG (i.e., the MAG was novel below that taxonomic rank).
B. Microbial community composition differences based on metagenomes from the Ganges, Amazon, Kalamas and Chattahoochee Rivers. The graph represents the principal coordinates analysis (PCoA) of the metagenomes based on Mash distances. The rivers are denoted by different colours, and the sampling locations along the Ganges River are denoted by different symbols (see Figure key).

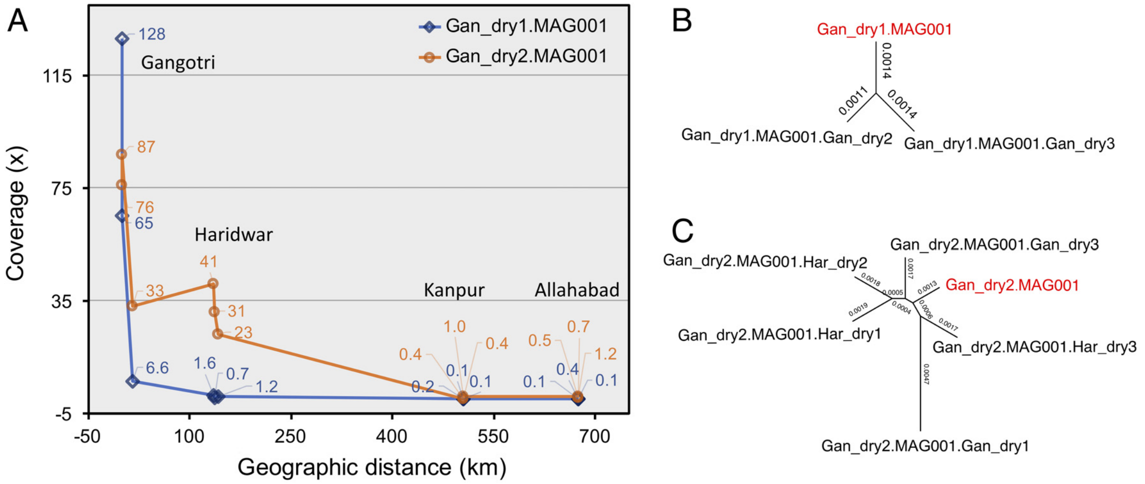

Fig. 5. Tracking of two upstream, abundant populations along the Ganges River.
A. Relative abundance (i.e., coverage) of two MAGs, that is, Gan_dry1.MAG001 and Gan_dry2.MAG001, based on the reads recruited from the 12 available metagenomes.B and C. Phylogenetic tree of reference MAGs, that is, Gan_dry1.MAG001 (B) and Gan_dry2.MAG001 (C), and consensus genomes assembled from each of the downstream samples. Only metagenomes with the coverage of the population higher than 5× were used in the analysis. Whole genome alignment was performed using Mugsy and phylogenetic relationships were inferred by Maximum likelihood as implemented in RAxML with optimization of substitution rates and GTRCAT model of 100 iterations for the rate of heterogeneity. Note the star-like phylogeny, with the possible exception of Gan_dry2.MAG001.Gan_dry1, which revealed that the population genomes were highly related to, and equal distant from, each other. [Color figure can be viewed at wileyonlinelibrary.com]

upstream samples Gan_dry1 and Gan_dry2 were tracked in downstream samples, along the Ganges River. Gan_dry1.MAG001 likely represented a novel genus related to the antimicrobial-resistant *Rheinheimera* sp. A13L strain (Gupta *et al*., 2011) (67.6% genome-aggregate average amino acid identity [AAI]) (Konstantinidis and Tiedje, 2005), and Gan_dry2.MAG001 was a close relative of an emerging nosocomial pathogen *Acinetobacter junii* NZ CP019041 (98.3% AAI) (Bansal *et al*., 2017). Both populations were detectable downstream, albeit with 100–1000 times lower abundances in some samples (Fig. 5A and Supporting Information Fig. S6). The downstream populations were clearly members/strains of the same species (e.g., ANI > 97%) as the reference MAGs based on phylogenetic analysis of their (assembled) whole genome sequences from the corresponding (individual) metagenomes (Fig. 5B and C).

Microbial diversity increased in the wet season along the Ganges River

Samples collected in the wet season presented a significantly higher dissimilarity among themselves based on Mash distances than those collected in the dry season (Fig. 6A), indicating that heavy, monsoon-associated rain increased beta-diversity along the Ganges River. [Metagenomic datasets from the upstream location in Gangotri were excluded from this analysis since these samples showed more distinct community composition compared with the remaining Ganges samples (Supporting Information Fig. S7A).] Consistently, the alpha-diversity of the samples collected from the wet season measured by Nonpareil was higher than those from the dry season (Supporting Information Table S3). While a significantly negative distance-decay of microbial community similarity over geographic distance (r = 0.83, p < 0.001) was revealed in the dry season, the samples from the wet season did not share this pattern (r = 0.37, p > 0.1) (Fig. 6B), presumably due to greater importance of variable (i.e. site-specific) allochthonous inputs during the rainy season. Consistent with these findings, environmental variables including turbidity, TDS, hardness, alkalinity, pH, COD, TC, IC, TOC, sulfate, nitrate and chloride significantly affected microbial community composition in the dry season (Supporting Information Fig. S7B), and the dissimilarity based on environmental variables significantly correlated (r = 0.54, p < 0.001) with the geographic variables (Latitude, Longitude and geographic distance) (Supporting Information Fig. S7C).

Clustering based on Morisita distances of OTUs abundance revealed consistent pattern with the Mash-based results (Fig. 6C and Supporting Information Fig. S7A). Further analysis of samples from Haridwar and Kanpur sites that represented both the wet and dry seasons, showed higher Morisita similarities of samples from the same season than from the same sites, indicating that seasonal effects were more pronounced than spatial ones. In terms of functional distribution, 370 subsystems based on SEED level 3 were significantly differentially abundant (adjusted p value < 0.01, negative binomial test) across sites and seasons in total (Supporting Information Table S5). Clustering based on significantly differentially abundant functions

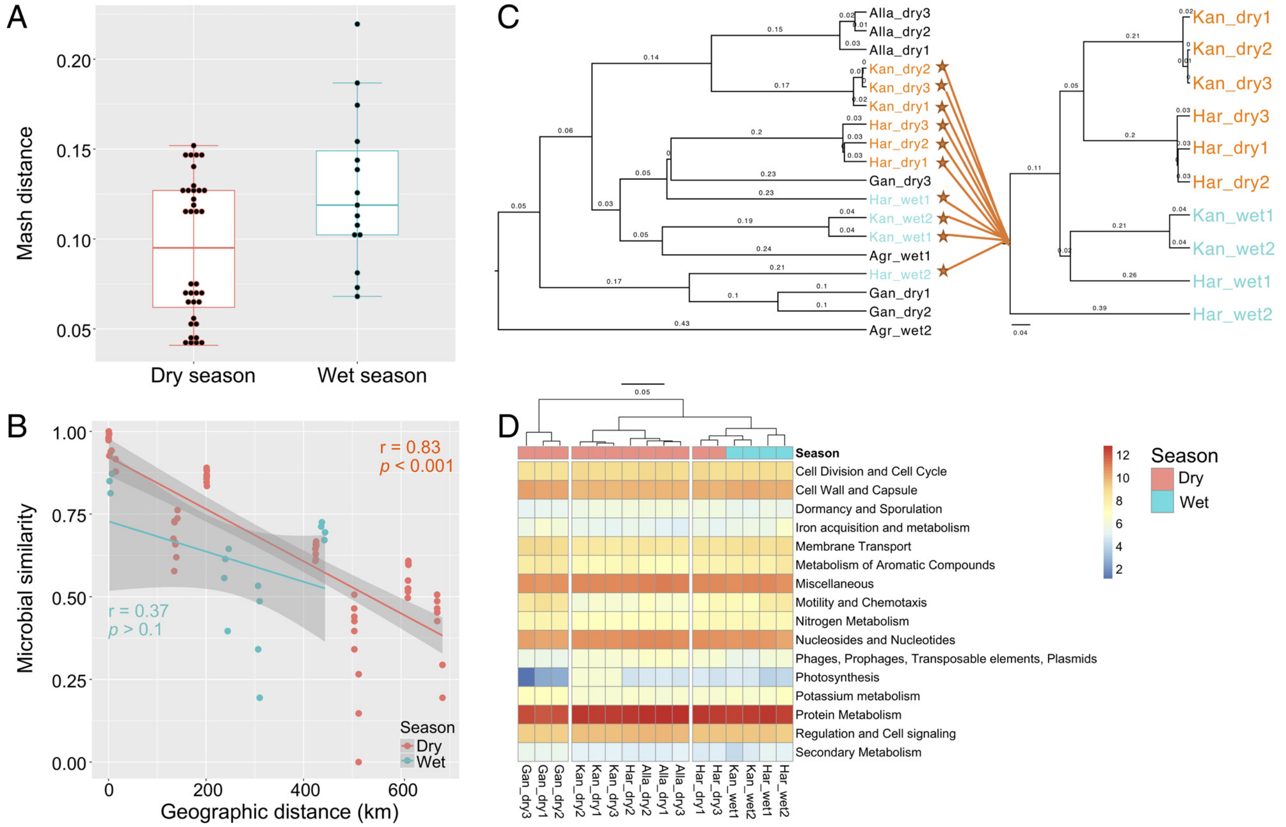

Fig. 6. The effect of the wet season on the Ganges microbial community composition.
A. Boxplot of Mash distances of metagenomes collected in the dry and wet seasons.
B. Distance-decay curve for microbial community similarity based on Mash distances.
C. Clustering of metagenomes based on Morisita distance of the 16S-based OTUs relative abundances. The smaller sub-tree only shows the samples from Haridwar and Kanpur sites that represented both the wet and dry seasons.
D. Heatmap of SEED subsystems (level 1) showing statistically significant differences in abundance between wet (blue) and dry (red) season samples (negative binomial test, adjusted *p* value < 0.01).

from pairwise comparisons also demonstrated the separation patterns of samples collected from different seasons. Lower abundance of genes associated with photosynthesis were observed in samples from the wet season, contrasting with higher abundances of nitrogen and aromatic compounds metabolism, and motility and chemotaxis associated genes (Fig. 6D).

Discussion

Intensive allochthonous inputs of terrestrial, sewage, sludge and human were detected in the Ganges River (Fig. 2A). Among them, sludge and sewage inputs most likely contributed to the high abundance of ARGs and HG sequences in the Ganges compared with freshwater ecosystems elsewhere (Fig. 3). Consistently, heavy municipal sewage and industrial discharges have often been dumped into the Ganges River (Sinha and Loganathan, 2015) and presumably sewage, which includes faecal microbiome material of human populations (Newton *et al*., 2015), accounted for the high abundance of HG sequences detected. In agreement with these interpretations, the limited treatment of sewage in India results in 80% of the sewage (untreated) to be released into the water system [https://phys.org/news/]. Faecal coliform counts have been also reported at high levels in rivers in India previously (Hamner *et al*., 2006). The detection of opportunistic pathogens such as HG microbes *Acinetobacter*, *Bacteroides* and *Prevotella* by our metagenomic approach (e.g., Supporting Information Fig. S3) was consistent with these previous results and revealed a potential threat for public health and for the millions of people who live along the banks of the River.

Recently, human gut microbiota has been reported as an important reservoir of ARGs (Feng *et al*., 2018). Thus, it is reasonable to hypothesize that the significantly high correlation of ARGs with HG sequences in the Ganges River (Supporting Information Fig. S5) resulted from ARGs carried by human gut associated microbes. Moreover, in India, due to the prevalent over-prescription and overuse of antibiotics in hospitals, clinics, households, animal farming and agri-industrial production, antibiotic-resistant pathogens are commonly detected in patients (Murki, 2009; Kotwani and Holloway, 2011; Wats and Sohal, 2013; Ganguly et al., 2011). The inappropriate use and disposal of antibiotics could also contribute to the widespread occurrence of ARGs in the environment (Guo *et al*., 2017). Consistent with these data, ARGs to about 90% of the commonly used antibiotics in India, including beta-lactams (cephalosporin, penicillin and carbapenem), fluoroquinolones, aminoglycosides, macrolides, tetracyclin and glycopeptide (vancomycin) (Wats and Sohal, 2013), were all detected in the Ganges River (Supporting Information Fig. S4). Corroboratively, a relatively high frequency of various antibiotic-resistant bacterial isolates, including extended-spectrum beta-lactamase producing Gram-negative bacteria, methicillin-resistant *Staphylococcus aureaus*, carbapenem-resistant *Pseudomonas*/*Acinetobacter*, vancomycin-resistant *Enterococci* and multidrug-resistant bacteria, have been previously reported in India (Murki, 2009). Collectively, our results showed that high abundance, especially compared with other freshwater ecosystems (see also below), and spreading of ARGs genes constitute another important public health risk in Ganges. Quantifying the associated risk for human health, including through irrigation of agricultural fields with water from the Ganges (Gorski *et al*., 2016; Cooley *et al*., 2018), should be subject of follow-up studies.

Somewhat unexpectedly, the upstream samples from Gangotri, which is much less populated by humans and thought to be a more pristine area compared with downstream areas, often had higher abundances of ARGs and HG sequences than some of the downstream samples (Supporting Information Fig. S2). Since the Gangotri samples were collected in May 2016, which coincided with massive numbers of visitors in this area during seasonal pilgrimages and bathing in the River (Ahammad *et al*., 2014), it is likely that the high signal of ARGs and HG sequences in these samples also originated from those visitors. Previous studies of the effect of pilgrimage-associated visits to the upper Ganges River has shown that seasonally higher $bla_{\mathrm{NDM-1}}$ (NDM-1 metallo-β-lactamase genes) levels resulted from the increased levels of faecal coliforms originating from the pilgrims in June (Ahammad *et al*., 2014). Consistently, we also noticed a relatively high number of pilgrims around Gangotri compared with the downstream locations during the sampling time. Noticeably, extremely higher turbidity (around 12–196 times) was detected in upstream samples in Gangotri (Supporting Information - Table S2), which possibly indicated the higher allochthonous inputs, mostly of soil origin through runoff, in the upper Ganges River. Thus, the higher signal of ARGs in the upstream sample was likely also attributable, at least in part, to microorganisms originating from soil, a well-known habitat for microbial antibiotic production.

In any case, as revealed and quantified by metagenomics, these variable allochthonous inputs in the Ganges River distinguish its microbial communities from their counterparts elsewhere, and dramatically alter both the taxonomic and functional composition in the river. While typical freshwater bacteria in rivers, including *Betaproteobacteria*, *Actinobacteria*, *Bacteroidetes* and *Verrucomicrobia* (Savio *et al*., 2015; Newton *et al*., 2011; Iliev *et al*., 2017) were also detected, the Ganges River was distinct by possessing a higher abundance of *Gammaproteobacteria*, which often have fast growth rates on high organic substrate conditions (copiotrophs) such as sewage lagoons (Newton *et al*., 2011). The higher allochthonous inputs rates of sludge/sewage altered microbial communities in Ganges River, especially in samples from the wet season and upstream. For example, a higher abundance of *Acinetobacter* and *Chromatiaceae* (*Gammaproteobacteria*) (Saunders *et al*., 2016; Bize *et al*., 2015; Wagner *et al*., 1994), *Comamonadaceae* (*Betaproteobacteria*) (Saunders *et al*., 2016), and the *Cytophaga-Flavobacterium* group (Liu *et al*., 2005) that are commonly found in sludge and sewage associated environments worldwide were detected in Ganges River in the wet season and upstream samples (Fig. 2B). The soil bacterium *Pseudomonas putida* (Nelson *et al*., 2002), the pathogen *Acinetobacter baumannii* (Antunes *et al*., 2014) and the antimicrobial-resistant *Theinheimera* sp. A13L (Gupta *et al*., 2011) were observed to be more abundant in upstream relative to downstream samples as well (Supporting Information Fig. S1).

In terms of gene content, the higher allochthonous inputs in the upstream and wet season decreased the abundance of photosynthesis genes, and led to a higher abundance of genes related to nitrogen and aromatic compound metabolism, mobility and chemotaxis, heavy metal (copper, arsenic and mercury) and antibiotic resistance (Fig. 6D and Supporting Information Table S5). These results were attributable, at least partly, to the high sewage and runoff inputs in the river and high usage of aromatic compounds, heavy metals and antibiotics in India (Kotwani and Holloway, 2011; Wats and Sohal, 2013; Singh and Pandey, 2014; Sharma *et al*., 2015), and were consistent with the phylogenetic diversity patterns described above.

At the individual population level, the two allochthonous populations, which are the most abundant in the upstream samples were trackable across the river after 1–2 days of (downstream) transport (> 200 km apart). Moreover, these sequence-discrete populations were abundant (> 5×) and robustly detected especially in the downstream samples less than 200 km. The intra-population sequence diversity measured by the identity of recruited reads to the reference MAG genome ($\mathrm{ANI_R}$

values) varied from 99.84% to 99.74%, indicating that the exact same population was identified at all sites based on the phylogenetic patterns of the consensus genome assembled from each sample or the recruited reads. Therefore, metagenomic read recruitment plots as described here hold great potential to reliably track, at high resolution, that is, the individual population level, the source of microbial taxa and their spatial and temporal patterns within an ecosystem such as the Ganges.

Consistent with expectations for the relative magnitude of run-off between wet versus dry seasons, increased abundances of HG sequences and OTUs related to sludge/sewage and terrestrial were detected in the Ganges River in the months with high levels of precipitation (e.g., July) (Fig. 2). Considering also that microbial communities in some of the source environments such as soil harbour a much greater diversity than aquatic communities (Crump *et al*., 2012), these substantial allochthonous inputs presumably accounted for the higher alpha-diversity of microbial communities in the wet season (Supporting Information Table S3). Although increased river flow due to rainfall has been reported to result in more community homogeneity in the other riverine systems (Carney *et al*., 2015; de Oliveira and Margis, 2015), a higher spatial heterogeneity was revealed in the Ganges River in the wet season (Fig. 6A). The Ganges River is one of the largest rivers in the world, and our sampling sites represented an approximately 700 km-long transect along the river, which is three to four times longer than the length of the rivers showing more spatial homogeneity during wet seasons. Thus, it is likely that ‘mass-effect’ of increased ‘riparian influence’ resulting from locally autochthonous inputs (Leibold *et al*., 2004; Crump *et al*., 2007) were more important in shaping microbial communities in the wet season in the Ganges relative to other (smaller) rivers.

While harbouring higher spatial heterogeneity, in the wet season, microbial communities showed no significant distance-decay patterns (Fig. 6B). Considering that dispersal-driven assembly mechanisms have been primarily detected only when beta-diversity is relatively low (Langenheder *et al*., 2012), environmental heterogeneity driven by ‘species sorting’ (Savio *et al*., 2015) could be the primarily dominant assembly force, instead of dispersal, for microbial communities in the wet season. In contrast, in the dry season, distance-decay of microbial community similarity was significant, indicating dispersal limitation was important in that season. In addition, previously identified environmental variables including alkalinity, pH, COD, nutrient availability (Savio *et al*., 2015; Staley *et al*., 2015; Zeglin, 2015; Jordaan and Bezuidenhout, 2016) and chloride concentrations (Stanish *et al*., 2016) also correlated with the Ganges microbial community beta-diversity patterns in the dry season (Supporting Information Fig. S7B). Nevertheless, after excluding geographic factors (latitude, longitude and geographic distance), environmental factors independently explained a small fraction (0.4%) of microbial community variation, about approximately 12-fold less compared with that explained by geographic factors independently (5%). It is thus reasonable to hypothesize that homogenizing physical and chemical conditions during the dry season across such a long riverine system is limited, as shown by the significant correlation of dissimilarities of environmental with geographic variables (Supporting Information Fig. S7C). Consistently, the majority (89%) was explained by co-effect of both geographic and environmental variables, reflecting dispersal limitations in microbial community assembly in the dry season.

The increased sampling volumes in the dry relative to the wet season (5 vs. 0.5 L) could have accounted, at least partly, for some of the results observed here in terms of community structure (Zinger *et al*., 2012), as volume size could affect diversity (Magurran, 2004), but it is not uniform among habitats (Prosser *et al*., 2007). However, we observed higher diversity in wet season samples, both in terms of alpha and beta diversity, which suggested that the effect of volume size was limited, if any. Consistent with the latter interpretation, low variability of community structure when sampling more than 50 ml of seawater (Ghiglione *et al*., 2005) and no significant differences in richness when sampling increasing volumes of water from 10 to 1000 ml (Dorigo *et al*., 2006) were shown previously based on fingerprint profiles.

In any case and to confirm these preliminary findings and interpretations about the microbial communities in Ganges River, which travels approximately 2700 km, more sampling sites/locations and different time points (months) will be needed. Clearly, the limited number of samples taken as part of our study did not allow for more robust conclusions to emerge with respect to several of the diversity patterns revealed compared with what was mentioned above. Nonetheless, the effects of anthropogenic inputs and wet season on microbial communities revealed by our analysis were significant. Moreover, our results suggest that metagenomics could offer robust means for reliable microbial source tracking purposes, even in challenging environments such as the Ganges River.

Experimental procedures

Sample collection and processing

Surface water samples were collected over an approximately 700 km long transect along the Ganges in India. In July 2015, in the wet season after the rains, two 0.5-L samples (biological replicates) were collected). In May

2016, in the dry season before the rains, three 5-L samples (biological replicates) were collected. The map of sampling sites was constructed using the ArcGIS platform, and overlaid with World Land cover 30 m BaseVue 2013 and World Population Estimate [available at: http://www.arcgis.com/] (Fig. 1). The water samples were filtered within 1–3 days after transport to the lab. All water samples were filtered using a peristaltic pump in the laboratory, first through a 2.7 μm pre-filter (142 mm) to remove large particles, followed by a 1.6 μm pore-size glass fibre filter (Geotech) to remove eukaryotic cells, and then collected onto a 0.2 μm Sterivex filter (Millipore) for 2015 samples or onto a 0.2 μm membrane filter (142 mm) for 2016 samples. The 0.2 μm filters were stored at −80°C until DNA extraction. DNA was extracted from the 0.2 μm filters using the cell lysis and organic extraction method as described previously (Oh *et al*., 2011) and detailed in Supproting Information. For the 2016 samples, physiochemical parameters including hardness, alkalinity, turbidity, pH, DO, BOD, COD, TDS, TC, IC, ammonium, nitrate, sulfate and chloride were measured using a HACH instrument.

Shotgun metagenomics sequencing, and sequence trimming, assembling and functional annotation

All 18 DNA libraries were prepared with the Illumina Nextera XT DNA library prep kit and sequenced on an Illumina Hiseq 2500 system for 300 cycles (paired end rapid run, 2 × 150 bp) at the High Throughput DNA Sequencing Core at the Georgia Institute of Technology. Metagenomic datasets have been deposited in National Center for Biotechnology Information (NCBI)'s Short Read Archive (SRA) database, under the bioproject PRJNA420715. NCBI SRA numbers for all datasets were provided in Supporting Information Table S1. Metagenome read quality checking, trimming and assembly as well as gene annotation of the resulting assembled contig sequences were performed as described in the Supporting Information Methods. Differentially abundant genes and subsystems between samples were determined with the DESeq package using the negative binomial model, adjusted for false discovery rate (Love *et al*., 2014). Significantly differentially abundant gene functions (adjust *p* value <0.01), based on SEED level 1 annotation were visualized by heatmap using the pheatmap R package (Kolde and Kolde, 2015).

16S rRNA gene-encoding read identification and taxonomic analysis

16S rRNA gene sequence (16S) fragments were extracted using Parallel-META (Su *et al*., 2012) from the metagenomics reads, and then processed for Operational Taxonomic Unit (OTU) picking, defined at the greater than 97% sequence threshold, and taxonomic identification with SILVA database (Pruesse *et al*., 2007) using QIIME 1.9.1 (Caporaso *et al*., 2010). Morisita distances (Wolda, 1981) were used to assess the microbial taxonomic compositional variation between samples based on the abundance of 16S-derived OTUs because of the independence of this metric from sample size and diversity. To validate the taxonomic classification of 16S-encoding metagenomic reads, MyTaxa with default parameters was also used to assess the bacterial and archaeal taxonomy of the assembled contigs (Luo *et al*., 2014). The input file to MyTaxa was the Blastp results of the query genes against the predicted proteins of all closed and draft prokaryotic genomes available in NCBI as of February 2015, with a minimum score of 60 bits and 80% of query protein sequence coverage by the alignment and only the top five hits per query, when available, being considered.

Detecting human gut microbes associated sequences and antibiotic resistance genes

Blastn search of predicted genes on assembled contigs against the Integrated Gene Catalogue (IGC) of genes of the human gut microbiome (HG Database) (Li *et al*., 2014) was performed to detect human gut microbes associated sequences (HG sequences). Only the best match with nucleotide identity ≥ 98% and reference sequence length coverage ≥ 50% by the alignment was considered as HG sequences. Taxonomic compositions of HG sequences were classified according to the HG Database. Antibiotic resistance genes (ARGs) were identified based on searches of predicted protein sequences on assembled contigs against the comprehensive antibiotic resistance database (CARD; April 2016 release) (McArthur *et al*., 2013) and a minimum cut-off for a match of 50% amino-acid identity and reference sequence length coverage of 50% by the alignment (only best matches were considered). Identified ARGs were annotated according to the categories available by CARD.

The abundance of HG sequences and ARGs in each metagenomic dataset was calculated by the number of reads mapping on each gene above the cut-off (identity ≥ 95% and query sequence coverage by the alignment ≥ 50%) using BLAT (Kent, 2002) normalized by the gene length. The abundance of RpoB genes was calculated by total reads that identified as RpoB genes by ROCker as described before (Orellana *et al*., 2016), normalized by the average length of the RpoB genes [the ROCker model is available at: http://enve-omics.ce.gatech.edu/rocker/models]. The genome equivalent, that is, how many cells of the total sampled encoded the HG sequences or ARGs, was estimated by the ratio of the abundance of the HG/ARG genes against that of the RpoB genes.

Population genome binning and relative in-situ abundance

Metagenome-assembled genomes (MAGs) were recovered using MaxBin Version 2.1.1 (Wu *et al*., 2014). Contigs longer than 500 bp from each metagenomic dataset and from coassembly of biological replicates were used for binning. Quality assessment of the resulting MAG sequences was performed by CheckM (Parks *et al*., 2015). Only the genomes with completeness ≥ 70% and contamination ≤ 10% were considered for further analysis. Taxonomic classification of obtained MAGs, including assessment of their taxonomic novelty, was performed by the Microbial Genomes Atlas (MiGA) webserver (Rodriguez-R et al., 2018a, b). The recovered MAGs were also compared with 1126 MAGs recovered from Chattahoochee River in the United States (freshwater ecosystem) [available in MiGA].

The two most abundant population genomes recovered from the Gangotri samples (Gan_dry1.MAG001 and Gan_dry2.MAG001), after a reassembly step to improve quality (see Supporting Information Methods for details) were tracked along the Ganges River. More specifically, the reads recruited from the 12 available metagenomics datasets originating from downstream samples in the dry season were mapped against these two MAGs using a BLAT search, and only the best match with alignment length ≥ 100 bp and 95% nucleotide identity was considered for coverage estimation using the enveomics.R package (Rodriguez-R and Konstantinidis, 2016). Phylogenetic analysis to test if the population in a metagenomic sample was the same strain as the reference MAG are detailed in the Supporting Information Methods.

Statistical analysis of alpha and beta community diversity patterns

Nonpareil, an algorithm that determines coverage based on the level of redundancy of the sequence reads of metagenomes (Rodriguez-R and Konstantinidis, 2014a, b), was used to estimate the abundance-weighted average coverage of the sampled microbial communities achieved by sequencing. The sample-to-sample sequence composition similarity was assessed by Mash distance (Ondov *et al*., 2016), and the resulting distance matrix was visualized by principal coordinate analysis (PCoA). The effects of environmental variables on beta-diversity patterns were first evaluated using a distance-based redundancy analysis coupled with analysis of variance (dbRDA-ANOVA) with 10 000 permutations based on the dissimilarity matrix of Mash distances; with only the significant variables included for further analysis. The effects of environmental variables, geographic distance (the distance of each site to the upstream location in Gangotri) and rainfall were subsequently summarized by nonmetric multidimensional scaling (NMDS). To assess the correlation between environmental variables and geographic variables (latitude, longitude and geographic distance), dissimilarity of environmental/geographic variables based on Euclidean distance was calculated and tested by Pearson correlation. The distance-decay analysis was applied to investigate the decrease in microbial community similarity (based on Mash distances) with geographic distance (sample-to-sample distance) in wet versus dry seasons in the Ganges River, and tested by Pearson correlation. These analyses were performed in R 3.2.3 with Vegan (Wagner, 2015) and ggplot2 (Hadley, 2015) packages.

Acknowledgements

This research was supported, in part, by the US National Science Foundation, award number DEB 1241046 (to KTK). We appreciate the support from the Brook Byers Institute for Sustainable Systems and the Hightower Chair at the Georgia Institute of Technology. The views and ideas expressed herein are solely those of the authors and do not represent the ideas of the funding agencies in any form.

References

Abia, A. L. K., Alisoltani, A., Keshri, J., and Ubomba-Jaswa, E. (2018) Metagenomic analysis of the bacterial communities and their functional profiles in water and sediments of the Apies River, South Africa, as a function of land use. *Sci Total Environ* **616**: 326–334.

Ahammad, Z. S., Sreekrishnan, T. R., Hands, C. L., Knapp, C. W., and Graham, D. W. (2014) Increased waterborne *bla*$_{NDM-1}$ resistance gene abundances associated with seasonal human pilgrimages to the upper Ganges river. *Environ Sci Technol* **48**: 3014–3020.

Allen, M. S., Welch, K. T., Prebyl, B. S., Baker, D. C., Meyers, A. J., and Sayler, G. S. (2004) Analysis and glycosyl composition of the exopolysaccharide isolated from the floc-forming wastewater bacterium *Thauera* sp. MZ1T. *Environ Microbiol* **6**: 780–790.

Antunes, L., Visca, P., and Towner, K. J. (2014) *Acinetobacter baumannii*: evolution of a global pathogen. *Pathog Dis* **71**: 292–301.

Bansal, P., Kaul, V., Easaw, S., and McGarry, T. (2017) Multi drug resistant *Acinetobacter junii* pneumonia is a ventilator dependent patient: a rare organism in critically sick adults. *Am J Respir Crit Care Med* **195**: A4066–A4066.

Bize, A., Cardona, L., Quemener, E. D.-L., Battimelli, A., Badalato, N., Bureau, C., *et al*. (2015) Shotgun metaproteomic profiling of biomimetic anaerobic digestion processes treating sewage sludge. *Proteomics* **15**: 3532–3543.

Buhtiani, R., Khanna, D. R., Kulkarni, D. B., and Ruhela, M. (2016) Assessment of Ganges river ecosystem at Haridwar, Uttarakhand, India with reference to water quality indices. *Appl Water Sci* **6**: 107–113.

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

Carney, R. L., Mitrovic, S. M., Jeffries, T., Westhorpe, D., Curlevski, N., and Seymour, J. R. (2015) River bacterioplankton community responses to a high inflow event. *Aquat Microb Ecol* **75**: 187–205.

Chao, Y., Ma, L., Yang, Y., Ju, F., Zhang, X. X., Wu, W. M., and Zhang, T. (2013) Metagenomic analysis reveals significant changes of microbial compositions and protective functions during drinking water treatment. *Sci Rep* **3**: 3550.

Cooley, M. B., Carychao, D., and Gorski, L. (2018) Optimized co-extraction and quantification of DNA from enteric pathogens in surface water samples near produce fields in California. *Front Microbiol* **9**: 448.

Crump, B. C., Adams, H. E., Hobbie, J. E., and Kling, G. W. (2007) Biogeography of bacterioplankton in lakes and streams of an arctic tundra catchment. *Ecology* **88**: 1365–1378.

Crump, B. C., Amaral-Zettler, L. A., and Kling, G. W. (2012) Microbial diversity in arctic freshwaters is structured by inoculation of microbes from soils. *ISME J* **6**: 1629–1639.

de Oliveira, L. F. V., and Margis, R. (2015) The source of the river as a nursery for microbial diversity. *PLoS One* **10**: e0120608.

Dorigo, U., Fontvieille, D., and Humbert, J. F. (2006) Spatial variability in the abundance and composition of the free-living bacterioplankton community in the pelagic zone of Lake Bourget (France). *FEMS Microbiol Ecol* **58**: 109–119.

Feng, J., Li, B., Jiang, X., Yang, Y., Wells, G. F., Zhang, T., and Li, X. (2018) Antibiotic resistome in a large-scale healthy human gut microbiota deciphered by metagenomic and network analyses. *Environ Microbiol* **20**: 355–368.

Ganguly, N. K., Arora, N. K., Chandy, S. J., Fairoze, M. N., Gill, J. P., Gupta, U., *et al*. (2011) Rationalizing antibiotic use to limit antibiotic resistance in India+. *Indian J Med Res* **134**: 281–294.

Ghai, R. (2011) Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS One* **6**: e23785.

Ghiglione J.F., Larcher M., and Lebaron P. (2005) Spatial and temporal scales of variation in bacterioplankton community structure in the NW Mediterranean Sea. *Aquat Microb Ecol* **40**: 229–240.

Gillings, M. R., Gaze, W. H., Pruden, A., Smalla, K., Tiedie, J. M., and Zhu, Y. G. (2015) Using the class 1 integron-integrase gene as a proxy for anthropogenic pollution. *ISME J* **9**: 1269–1279.

Gorski, L., Parker, C. T., Liang, A. S., Walker, S., and Romanolo, K. F. (2016) The majority of genotype of the virulence gene *inlA* are intact among natural watershed isolates of *Listeria monocytogenes* from the Central California coast. *PLoS One* **11**: e0167566.

Guo, J., Li, J., Chen, H., Bond, P. L., and Yuan, Z. (2017) Metagenomic analysis reveals wastewater treatment plants as hotspots of antibiotic resistance genes and mobile genetic elements. *Water Res* **123**: 468–478.

Gupta, H. K., Gupta, R. D., Singh, A., Chauhan, N. S., and Sharma, R. (2011) Genome sequence of *Rheinheimera* sp. strain A13L, isolated from Pangong Lake, India. *J Bacteriol* **193**: 5873–5874.

Hadley, W. R ggplot2 package: an implementation of the grammar of graphics. 2015.

Hamner, S., Tripathi, A., Mishra, R. K., Bouskill, N., Broadawat, S. C., Pyle, B. H., *et al*. (2006) The role of water use patterns and sewage pollution in incidence of water-borne/enteric diseases along the Ganges river in Varanasi, India. *Int J Environ Health Res* **16**: 113–132.

Hamner, S., Broadaway, S. C., Mishra, V. B., Tripathi, A., Mishra, R. K., Pulcini, E., *et al*. (2007) Isolation of potentially pathogenic *Escherichia coli* O157: H7 from the Ganges River. *Appl Environ Microbiol* **73**: 2369–2372.

Iliev, I., Yahubyan, G., Marhova, M., Apostolova, E., Gozmanova, M., Gecheva, G., *et al*. (2017) Metagenomic profiling of the microbial freshwater communities in two Bulgarian reservoirs. *J Basic Microbiol* **57**: 669–679.

Jordaan, K., and Bezuidenhout, C. (2016) Bacterial community composition of an urban river in the north West Province, South Africa, in relation to physico-chemical water quality. *Environ Sci Pollut Res* **23**: 5868–5880.

Kent, W. J. (2002) BLAT-the BLAST-like alignment tool. *Genome* **12**: 656–664.

Kolde, R., Kolde, M.R. (2015) Package 'pheatmap'.

Konstantinidis, K. T., and Tiedje, J. M. (2005) Genomic insights that advance the species definition for prokaryotes. *Proc Natl Acad Sci U S A* **102**: 2567–2572.

Kotwani, A., and Holloway, K. (2011) Trends in antibiotic use among outpatients in New Delhi, India. *BMC Infect Dis* **11**: 99.

Langenheder, S., Berga, M., Ostman, O., and Szekely, A. J. (2012) Temporal variation of β-diversity and assembly mechanisms in a bacterial metacommunity. *ISME J* **6**: 1107–1114.

Leibold, M. A., Holyoak, M., Mouquet, N., Amarasekare, P., Chase, J. M., Hoopes, M. F., *et al*. (2004) The metacommunity concept: a framework for multi-scale community ecology. *Ecol Lett* **7**: 601–613.

Li, J., Jia, H., Cai, X., Zhong, H., Feng, Q., Sunagawa, S., *et al*. (2014) An integrated catalog of reference genes in the human gut microbiome. *Nat Biotechnol* **32**: 834–841.

Liu, Y., Zhang, T., and Fang, H. H. (2005) Microbial community analysis and performance of a phosphate-removing activated sludge. *Bioresour Technol* **96**: 1205–1214.

Love, M. I., Huber, W., and Anders, S. (2014) Moderated estimation of fold change and dispersion for RNA-seq data with DESeq2. *Genome Biol* **15**: 550.

Luo, C., Rodriguez-R, L. M., and Konstantinidis, K. T. (2014) MyTaxa: an advanced taxonomic classifier for genomic and metagenomic sequences. *Nucleic Acids Res* **42**: e73.

Magurran, A. E. (2004) *Measuring Biological Diversity*. Oxford: Blackwell Publishing.

Maixner, F., Wagner, M., Lucker, S., Pelletier, E., Schmitz-Esser, S., Hace, K., *et al*. (2008) Environmental genomics reveals a functional chlorite dismutase in the nitrite - oxidizing bacterium 'Candidatus Nitrospira defluvii'. *Environ Microbiol* **10**: 3043–3056.

Marti, E., Jofre, J., and Balcazar, J. L. (2013) Prevalence of antibiotic resistance genes and bacterial community composition in a river influenced by a wastewater treatment plant. *PLoS One* **8**: e78906.

McArthur, A. G., Waglechner, N., Nizam, F., Yan, A., Azad, M. A., Baylay, A. J., *et al*. (2013) The comprehensive antibiotic resistance database. *Antimicrob Agents Chemother* **57**: 3348–3357.

Meziti, A., Tsementzi, D., Kormas, K. A., Karayanni, H., and Konstantinidis, K. T. (2016) Anthropogenic effects on bacterial diversity and function along a river to estuary gradient in Northwest Greece revealed by metagenomics. *Environ Microbiol* **18**: 4640–4652.

Murki, S. (2009) Antibiotic usage and microbial resistance: Indian scenario. *Neonatology* **23**: 53–56.

Namrata, S. (2010) Physicochemical properties of polluted water of river ganga at Varanasi. *IJEE* **1**: 823–832.

Nelson, K. E., Weinel, C., Paulsen, I. T., Dodson, R. J., Hilbert, H., Martins dos Santos, V. A. P., *et al*. (2002) Complete genome sequence and comparative analysis of the metabolically versatile pseudomonas putida KT2440. *Environ Microbiol* **4**: 799–808.

Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D., and Bertilsson, S. (2011) A guide to the natural history of freshwater lake bacteria. *Microbiol Mol Biol Rev* **75**: 14–49.

Newton, R. J., McLellan, S. L., Dila, D. K., Vineis, J. H., Morrison, H. G., Eren, A. M., *et al*. (2015) Sewage reflects the microbiomes of human populations. *MBio* **6**: e02574–e02514.

Oh, S., Caro-Quintero, A., Tsementzi, D., DeLeon-Rodriguez, N., Luo, C., Poretsky, R., and Konstantinidis, K. T. (2011) Metagenomic insights into the evolution, function, and complexity of the planktonic microbial community of Lake Lanier, a temperate freshwater ecosystem. *Appl Environ Microbiol* **77**: 6000–6011.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S., *et al*. (2016) Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biol* **17**: 132.

Orellana, L. H., Rodriguez-R, L. M., and Konstantinidis, K. T. (2016) ROCker: accurate detection and quantification of target genes in short-read metagenomic data sets by modeling sliding-window bitscores. *Nucleic Acids Res* **45**: e14.

Parks, D. H., Imelfort, M., Skeneerton, C. T., Hugenholtz, P., and Tyson, G. W. (2015) CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* **25**: 1043–1055.

Prosser, J. I., Bohannan, B. J. M., Curtis, T. P., Ellis, R. J., Firestone, M. K., Freckleton, R. P., *et al*. (2007) Essay the role of ecological theory in microbial ecology. *Nat Rev Microbiol* **5**: 384–392.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glockner, F. O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* **35**: 7188–7196.

Rodriguez-R, L. M., and Konstantinidis, K. T. (2014a) Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *BMC Bioinform* **30**: 629–635.

Rodriguez-R, L. M., and Konstantinidis, K. T. (2014b) Estimating coverage in metagenomic data sets and why it matters. *ISME J* **8**: 2349–2351.

Rodriguez-R, L. M., and Konstantinidis, K. T. (2016) The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints* **4**: e1900v1.

Rodriguez-R, L. M., Gunturu, S., Harvey, W. T., Rossello-Mora, R., Tiedje, J. M., Cole, J. R., *et al*. (2018a) The microbial genomes atlas (MiGA) webserver: taxonomic and gene diversity analysis of archaea and bacteria at the whole genome level. *Nucleic Acids Res* **46**: 282–288.

Rodriguez-R, L. M., Gunturu, S., Tiedje, J. M., Cole, J. R., and Konstantinidis, K. T. (2018b) Nonpareil 3: fast estimation of metagenomic coverage and sequence diversity. *Noval Syst Biol Techniq* **3**: e00039-18.

Sangwan, P., Kovac, S., Davis, K. E. R., Sait, M., and Janssen, P. H. (2005) Detection and cultivation of soil *Verrucomicrobia*. *Appl Environ Microbiol* **71**: 8402–8410.

Saunders, A. M., Albertsen, M., Vollertsen, J., and Nielsen, P. H. (2016) The activated sludge ecosystem contains a core community of abundant organisms. *ISME J* **10**: 11–20.

Savio, D., Sinclair, L., Ijaz, U. Z., Parajka, J., Reischer, G. H., Stadler, P., *et al*. (2015) Bacterial diversity along a 2600 km river continuum. *Environ Microbiol* **17**: 4994–5007.

Sharma, B. M., Bharat, G. K., Tayal, S., Nizzetto, L., Cupr, P., and Larssen, T. (2014) Environment and human exposure to persistent organic pollutants (POPs) in India: a systematic review of recent and historical data. *Environ Int* **66**: 48–64.

Sharma, B. M., Nizzetto, L., Bharat, G. K., Tayal, S., Melymuk, L., Sanka, O., *et al*. (2015) Melting Himalayan glaciers contaminated by legacy atmospheric depositions are important sources of PCBs and high-molecular-weight PAHs for the Ganges floodplain during dry periods. *Environ Pollut* **206**: 588–596.

Singh, A. V., and Pandey, J. (2014) Heavy metals in the midstream of the Ganges River: spatio-temporal trends in a seasonally dry tropical region (India). *Water Int* **39**: 504–516.

Sinha, R.K., and Loganathan, B.G. (2015) Ganges river contamination: a review. In *Water Challenges and Solutions on a Global Scale, ACS Symposium Series. American Chemical Society*, pp. 129–159.

Sood, A., Singh, K. D., Pandey, P., and Sharma, S. (2008) Assessment of bacterial indicators and physicochemical parameters to investigate pollution status of Gangetic river system of Uttarakhand (India). *Ecol Indic* **8**: 709–717.

Staley, C., Gould, T. J., Wang, P., Philips, J., Cotner, J. B., and Sadowsky, M. (2015) Species sorting and seasonal dynamics primarily shape bacterial communities in the upper Mississippi River. *Sci Total Environ* **505**: 435–445.

Stanish, L. F., Hull, N. M., Robertson, C. E., Harris, J. K., Stevens, M. J., Spear, J. R., and Pace, N. R. (2016)

Factors influencing bacterial diversity and community composition in municipal drinking waters in the Ohio River basin, USA. *PLoS One* **11**: e0157966.

Stover, C. K., Pham, X. Q., Erwin, A. L., Mizoguchi, S. D., Warrener, P., Hickey, M. J., *et al*. (2000) Complete genome sequence of *Pseudomonas aeruginosa* PAO1, an opportunistic pathogen. *Nature* **406**: 959–964.

Su, X., Xu, J., and Ning, K. (2012) Parallel-META: efficient metagenomic data analysis based on high-performance computation. *BMC Syst Biol* **6**: S16.

Tyagi, V. K., Bhatia, A., Gaur, R. Z., Khan, A. A., Ali, M., Khursheed, A., *et al*. (2013) Impairment in water quality of Ganges river and consequential health risks on account of mass ritualistic bathing. *Desalin Water Treat* **51**: 2121–2129.

Wagner, H. (2015) *Vegan: Community Ecology Package. R Package Version 2.0.*

Wagner, M., Erhart, R., Manz, W., Amann, R., Lemmer, H., Wedi, D., and Schleifer, K. H. (1994) Development of an rRNA-targeted oligonucleotide probe specific for the genus *Acinetobacter* and its application for in situ monitoring in activated sludge. *Appl Environ Microbiol* **60**: 792–800.

Wats, A., and Sohal, S. (2013) Trends of antibiotic use among the indoor patients of medicine and pediatric Ward at a tertiary care hospital. *Int J Sci Res* **4**: 229–232.

Wolda, H. (1981) Similarity indices, sample size and diversity. *Oecologia* **50**: 296–302.

Wu, Y. W., Tang, Y. H., Tringe, S. G., Simmons, B. A., and Singer, S. W. (2014) MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome* **2**: 26.

Zeglin, L. H. (2015) Stream microbial diversity in response to environmental changes: review and synthesis of existing research. *Front Microbiol* **6**: 454.

Zinger, L., Gobet, A., and Pommier, T. (2012) Two decades of describing the unseen majority of aquatic microbial diversity. *Mol Ecol* **21**: 1878–1896.

Supporting Information

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

Fig. S1. Species-level community composition and abundance of taxa in the Ganges River. Taxonomy was based on MyTaxa analysis of assembled contigs and relative abundance was determined based on the mapping of metagenomic reads against the contigs using BLAT. Only the top 30 most abundant species are shown.

Fig. S2. The genome equivalents (average copies per cell) of (A) human gut microbiome associated and (B) antibiotic resistance gene sequences in samples from the Ganges, Amazon, Kalamas and Chattahoochee Rivers.

Fig. S3. Taxonomic classification of human gut associated microbes in the Ganges River at the genus level. Blastn searches of predicted genes on assembled contigs against the Integrated Gene Catalogue (IGC) of genes of the human gut microbiome (HG Database) were performed to detect human gut microbes associated sequences (HG sequences). Only the best matches with nucleotide identity ≥98% and reference sequence length coverage ≥50% by the alignment were considered HG sequences. Taxonomic compositions of the HG sequences were classified according to the HG Database.

Fig. S4. Functional annotation of the antibiotic resistance genes detected in the Ganges River. Antibiotic resistance genes (ARGs) were identified based on searches of predicted protein sequences on assembled contigs against the comprehensive antibiotic resistance database (CARD; April 2016 release) and a minimum cut-off for a match of 50% amino-acid identity and reference sequence length coverage of 50% by the alignment (only best matches were considered). Identified ARGs were annotated according to the categories available by CARD.

Fig. S5. Pearson correlation of the relative abundance of human gut microbiome associated and antibiotic resistance gene sequences in Ganges River. The fit-line in blue indicates significant correlation ($r = 0.63$, $p < 0.01$).

Fig. S6. An example of a recruitment plot of reads recruited from the metagenomics datasets originating from downstream samples (Gan_dry2) mapped against the reference MAG (Gan_dry1.MAG001). For further details, please see main text.

Fig. S7. (A) Two-dimensional NMDS plots based on Mash distances of metagenomics datasets collected from the Ganges River in wet and dry seasons. The correlation of the ordination scores with rainfall and geographic distance is displayed as grey vectors. (B) Two-dimensional NMDS plots based on Mash distances of metagenomics datasets collected from the Ganges River in the dry season. The correlation of the ordination scores with environmental variables is displayed as grey vectors. (C) Pearson correlation of the dissimilarity of environmental variables and geographic variables (Latitude, Longitude and geographic distance). The environmental variables displayed in Fig. S7 were first evaluated using a distance-based redundancy analysis coupled with analysis of variance (dbRDA-ANOVA) with 10 000 permutations based on the dissimilarity matrix of Mash distances, and only the significant variables were selected for the NMDS analysis. The geographic distance is estimated by the distance of each site to the upstream location in Gangotri. Euclidean distances were calculated to estimate the dissimilarity of environmental and geographic variables. Samples collected from wet season included Har_wet1 and Har_wet2 (light blue), Kan_wet1 and Kan_wet2 (light green), Agr_wet1 and Agr_wet2 (pink). Samples collected from dry season included Gan_dry1, Gan_dry2 and Gan_dry3 (red), Har_dry1, Har_dry2 and Har_dry3 (blue), Kan_dry1, Kan_dry2 and Kan_dry3 (green), Alla_dry1, Alla_dry2 and Alla_dry3 (black).

Table S1. Information of samples included in this study.

Table S2. Physicochemical parameters.

Table S3. Metagenomic dataset statistics.

Table S4. Information of MAGs assembled from 18 metagenomic datasets.

Table S5. Normalized abundance of significantly ($p < 0.01$) differentially abundant subsystems at SEED level 3 over location and sampling year among the Ganges metagenomes.