Title: Bacterial Community Fingerprints as Bioindicators of Watershed Land Cover

Authors: Jacob Gareis*, Osvaldo Martinez, Silas Bergen, Dylan Blumentritt, Carter Olsen

* Corresponding Author

Affiliations:

1| Department of Biology, Winona State University, Winona, MN, United States of America

2| Department of Geoscience, Winona State University, Winona, MN, United States of America

3| Department of Mathematics & Statistics, Winona State University, Winona, MN, United States of America

4| Department of Statistics, Virginia Tech, Blacksburg, VA, United States of America

Title: Bacterial Community Fingerprints as Bioindicators of Watershed Land Cover

Abstract:

Large river systems such as the Yamuna and the Mississippi are ecological and cultural lifelines and understanding the factors that govern their microbial communities is central to monitoring their health and resilience. Because bacteria are among the most metabolically diverse organisms on Earth and respond rapidly to environmental change, we hypothesized that riverine bacterial communities could serve as sensitive bioindicators of broad-scale watershed conditions. Using surface-water 16S rRNA gene amplicon data from eleven sites spanning the Upper Yamuna (India) and the Upper Mississippi (USA), we used k-means clustering to identify five distinct bacterial community types that corresponded to the sites' geographic regions. Similarity percentage analysis identified 29 taxa, each contributing at least 1% to the average Bray-Curtis dissimilarity between clusters. The pollutant-tolerant genera Arcobacter and Geobacter were the largest drivers of dissimilarity, reaching high relative abundances at the heavily urbanized New Delhi sites and remaining near-absent elsewhere. To relate these taxa to surrounding land use, we correlated their relative abundances against eight land-cover types at three nested spatial regions. The correlations depended strongly on buffer size. Within 5 kilometers of the sampling sites no association remained significant after multiple-testing correction, whereas twenty were significant at a 50-kilometer buffer and eighteen for the whole upstream watershed. Sorted by the land-cover class each correlated with, most of the taxa fell into one of four groups: one associated with urban and built cover, a second with snow and ice, a third with open water, and a fourth with rangeland and tree cover. Together, these results indicate that a small subset of taxa may be able to serve as a compositional "fingerprint" of watershed conditions, and that the relevant spatial scale for that fingerprint is the broader watershed as opposed to the very local area.

Introduction:

In streams and rivers, heterotrophic bacteria, archaea, and fungi break down much of the organic matter that arrives and carry out the reactions that cycle nutrients (Findlay 2010). The microbes doing that work limit eutrophication, degrade toxins, and supply much of the food that sustains the rest of the food web (Hilderbrand et al. 2023).

Running waters were long treated as pipes carrying carbon from land to sea, but they are better understood as biogeochemical reactors that rank among the most heterotrophic ecosystems on Earth, respiring more carbon than they produce and running on organic matter washed in from the surrounding land (Battin et al. 2023).

A whole watershed drains into a single channel, and flow runs one way, so the consequences of activity anywhere in the network are carried downstream and accumulate (Hilderbrand et al. 2023). Water collected at one point in a river therefore reflects the entire area drained above it. Everything the water carries, bacteria included, was gathered somewhere along the way. Four properties make the bacteria carried in that water particularly useful bioindicators of upstream conditions. First, bacteria are far more metabolically diverse than the macroinvertebrates and fish used in conventional biomonitoring. Bacteria break down a wide range of organic compounds and can respire using many chemicals other than oxygen. Bacteria also carry out the major transformations of nitrogen in streams, including nitrogen fixation and denitrification, neither of which any animal performs (Findlay 2010). Each of these processes requires a particular supply of chemicals, and the water delivers those chemicals from the whole area drained above the site, so the bacteria that thrive there reflect conditions throughout the watershed. Second, bacteria are small and have a high ratio of surface area to volume, which leaves them responsive to nutrients and pollutants at very low concentrations (Fontaine et al. 2023). Third, bacteria reproduce quickly and whole communities turn over rapidly, so river bacteria reorganize after a disturbance faster than the macroinvertebrate and fish communities used in conventional biomonitoring (Sagova-Mareckova et al. 2021). Fourth, macroinvertebrate and fish communities in heavily modified streams shrink to a few tolerant species, whereas bacterial diversity can stay high, so bacterial communities still distinguish among sites where macroinvertebrates and fish no longer do (Hilderbrand et al. 2020). Together, these four properties have motivated a growing effort to use bacterial community composition as an indicator of watershed condition (Ma et al. 2022; Fontaine et al. 2023).

Empirical support for the idea that bacteria can serve as bioindicators has grown substantially in recent years, and that support runs from water chemistry through pollution status to land use itself. A year-long survey of rivers in southwestern British

Columbia found that surface-water bacterial communities grouped by water chemistry rather than by the watershed the sample came from. Agricultural sites carried a consistent bacterial signature that set them apart from unaffected sites, whereas urban sites were more variable and grouped with either agricultural or unaffected sites depending on their nutrient concentrations (Van Rossum et al. 2015).

Along 2,600 kilometers of the Danube River, bacterial community composition at upstream sites predicted the biological conditions downstream. These conditions were measured in two ways, with chlorophyll a as a proxy for eutrophication and the macroinvertebrate-based Saprobic Index as a measure of organic pollution (Fontaine et al. 2023). Land use itself can be recovered as well. In New Zealand, stream biofilm communities predicted the land-use type of the upstream watershed with roughly 65% accuracy across approximately 1,000 kilometers (Hermans et al. 2024), and soil bacterial communities from the same region classified sites into four land-use categories with 85% accuracy (Hermans et al. 2020).

Stream bacterial communities differ in composition according to the surrounding land use. In four Panamanian streams draining mature forest, secondary forest, silvopasture, and cattle pasture, the forested streams held higher phylogenetic diversity while the cattle pasture stream had lower diversity and a community unlike the others. The silvopasture stream resembled the forested streams during wet seasons and the pasture stream during dry ones (Chavarria et al. 2021). Much of the influence of land use is indirect. Across boreal streams in Finland, agricultural and forested land-use appeared most strongly in conductivity, pH, and phosphorus concentration, and those local conditions in turn shaped which taxa were present (Weigel et al. 2023).

Composition is not the only aspect of the community that responds to environmental conditions, however. In streams of southwestern China, benthic biofilms at sites downstream of developed land held more nutrients and organic carbon than biofilms at forested sites upstream. Communities that differed more in composition also differed more in functional genes, and the downstream communities showed stronger denitrification and carbon metabolism (Qu et al. 2017). In a study of the West Run Watershed, a mixed-land-use system in West Virginia, researchers sampled streambank

sediment every other week from mid-May to mid-September at six sites that differed in how much of the land draining to each site lay downstream of agriculture and urban development. The proportion of land under that development was positively correlated with how much the community changed over the season (adjusted $R2 = 0.65$), so upstream development registered as instability rather than as a consistently different composition. Stream sites at more developed points in the river also had a metabolic quotient 20 to 50% higher than the forested reference site, meaning their microorganisms respired more carbon per unit of biomass and converted less of it into new growth (Dang et al. 2021).

How strongly land use explains bacterial composition depends on the buffer size over which land use is measured. Across urban, agricultural, and natural watersheds in China, land-use explained bacterial composition best at different buffer sizes depending on the watershed type and season. In the dry season, the strongest association between land-use and composition was a 100-meter riparian buffer in the urban watershed and at a 500-meter buffer in the agricultural one (She et al. 2024). Bacterial composition is also not a simple function of how much land is developed. In an urbanizing watershed in southern China, the way developed land influenced sediment communities more non-linearly then the proportion of developed land was distributed along the river produced non-linear shifts in sediment communities, and land-use within 300 meters of the channel was significantly associated with sediment community composition (Zhang et al. 2025).

Land cover influences river bacterial communities through several mechanisms. We organize those mechanisms here into three pathways, namely dissolved chemistry, particles, and the transport of bacteria themselves.

Dissolved chemistry is the first pathway. Agriculture and urban development rase the nitrogen, phosphorus, and organic carbon loads reaching streams (Qu et al. 2017; Secchi and McDonald 2019). Land application of poultry manure adds a pulse of labile, tryptophan-like dissolved organic matter to field runoff, and Singh et al. (2014) measured that pulse declining over the one to three weeks following application. Industrial land can also deliver metals. Along the Longsha River in China, Pan et al.

(2022) traced sediment heavy metals mainly to industrial land and highways and reported that metal-tolerant genera dominated the sediment near industrial land, while genera involved in carbon, nitrogen, and sulfur cycling were less abundant there.

Particles are the second pathway. Livestock grazing increases the suspended sediment, faecal material, and turbidity delivered to streams (Agouridis et al. 2005). Suspended sediment offers particle-associated bacteria more surface to colonize, and higher turbidity reduces the light available to photosynthetic organisms present.

The movement of bacteria from land into the water is the third pathway. Cailon et al. (2021) found that roughly 70% of the bacterial operational taxonomic units in pre-alpine streams and their biofilms were also present in soil water, and that bacterial diversity in headwater streams was significantly higher at high flow than at low flow. High flow therefore delivers recurring inoculation of soil bacteria to the stream. How the land is used determines the composition of that soil source. In the Yellow River floodplain, Dong et al. (2023) found that converting natural shrubland to farmland, grassland, or forest significantly changed the structure of the floodplain soil microbial community.

Dissolved chemistry, particles, and bacterial transport differ in how far they reach. Dissolved nutrients and metals travel in solution and can arrive from far upstream. Sediment particles settle as they move downstream, so particle-based influence is weighted closer to the source. Bacteria enter the water primarily during storms and snowmelt, and mostly from the parts of the watershed that generate runoff. No single buffer size should therefore be expected to capture all three pathways.

The effects of land use are not confined to the area immediately around a sampling point. Ahlinder et al. (2024) surveyed five Swedish lakes and their inflows and found that the presence of buildings, farms, and fields was associated with the abundance of roughly a third of the most common inflow taxa. The inflows also carried more iron, aluminum, and humic color than the lake water, and about a fifth of the most abundant taxa present in the lake were estimated to have come from the inflows.

Most studies nonetheless select one delineation in advance, either the whole upstream watershed or a buffer of chosen width or radius, and then ask whether land cover inside that delineation explains the community structure. Fewer studies have compared delineations against one another within a single dataset, and those that have compared very different delineation scales. Zhang et al. (2025) tested buffers from 100 to 500 meters along the Liuxi River in China and found the strongest associations with sediment communities at 300 meters. She et al. (2024) tested a 100-meter riparian buffer, a 500-meter riparian buffer, and the sub-watershed, meaning the drainage area above each sampling site. They found the correlations at the 500-meter and the sub-watershed to be of similar strength. Neither study examined the gap between a 500-meter buffer and the whole drainage area above the site, which is where the delineations compared in this study fall.

We compare land cover across those delineations in the Yamuna River (India) and the Mississippi River (USA). Both rivers drain large watersheds under heavy human pressure, though the pressure takes a different form in each. On the Yamuna River, the pressure is concentrated with 2% of its watershed lying inside the National Capital Territory of Delhi, and that 48-kilometer stretch collects 79% of the river's total pollution (Sharma et al. 2024). On the Mississippi River the pressure is dispersed. Fertilizer and manure run off farmland across the watershed, and Secchi and McDonald (2019) identify agricultural runoff as the main driver of the poor water quality in the Mississippi River.

Our earlier survey of eleven sites found similar bacterial community composition at most sites on both the Yamuna River and the Mississippi River (Martinez et al. 2024). The exceptions were the sites nearest the glacier-fed Yamuna River headwaters, which had elevated diversity compared to the other sites, and the two New Delhi sites, which were enriched with Arcobacter and Geobacter, two pollutant-tolerant genera.

Building on these findings, we aim to identify a small set of bacterial taxa whose relative abundances relate to land cover in the contributing watershed, and to determine the

delineation over which that relationship is strongest. We refer to the relative abundances of that set of taxa as a fingerprint of watershed condition. We expected urban-influenced watersheds to have a unique fingerprint with higher abundances of pollutant-tolerant taxa. We first used k-means clustering to define bacterial community types and similarity percentage (SIMPER) analysis to identify the taxa responsible for the differences between those community types. That step reduced 13,466 amplicon sequence variants (ASVs), the unique 16S rRNA gene sequences obtained from the sequencing data to a short set of candidate indicator taxa. We then correlated the relative abundances of those taxa against eight land-cover classes measured over three nested delineations, namely a 5-kilometer buffer, a 50-kilometer buffer, and the whole upstream watershed.

Materials and Methods:

Sample Collection and Filtration

We previously published an initial comparison of the riverine microbial populations at eleven sites across the Upper Yamuna and Upper Mississippi River systems between July 12th and August 2nd of 2019 (Table 1, Figure 1; Martinez et al. 2024). Surface water sample collection, sample DNA extraction, 16rRNA DNA sequencing and taxonomic assignment was previously described (reference). The final dataset included 13,466 unique ASVs across the 36 sequenced libraries. Raw sequences were deposited in the NCBI Sequence Read Archive under BioProject accession PRJNA1077596**.**

Data Preprocessing

All analyses were conducted in R version 4.4.3 (R Core Team 2025). Sequence counts were averaged across triplicates within sampling site. The ASV table was then collapsed into a site-by-taxon matrix by aggregating counts at the species level, yielding 1,423 taxa after removal of entries with zero total abundance across all sites. Relative abundances were obtained by dividing each count by the total sequencing depth for that site.

Clustering of Microbial Communities

k-means clustering (Hartigan and Wong 1979) was performed on the site-by-taxon relative abundance matrix to identify bacterial community types. The number of clusters was selected from an elbow plot of the within-cluster sum of squares.

SIMPER Analysis and ASV Selection

Similarity percentage (SIMPER) analysis (Clarke 1993) was applied to the count matrix using the simper() function of the *vegan* package (Oksanen et al. 2025) to identify the taxa contributing most to between-cluster dissimilarity. For each pairwise cluster comparison, taxa were ranked by their average contribution to Bray-Curtis dissimilarity. Taxa whose mean contribution across comparisons was at least 1% were retained. We then created plots to investigate the relative abundance of retained taxa across and within each of the five clusters.

Watershed Delineation and Land use Variables

Watershed boundaries were delineated in ArcGIS Pro (Esri Inc. 2024). Digital elevation data were obtained from the HydroSHEDS database (Lehner et al. 2008). Sinks were removed with the Fill tool to ensure continuous flow paths, flow direction was computed with the D8 algorithm, and flow accumulation was used to identify drainage patterns. Pour points were placed at the sampling locations and the Watershed tool was used to delineate the upstream watershed of each site.

Land-cover data came from the 2019 ESRI Sentinel-2 10 m Global Land Cover product (Karra et al. 2021), accessed through the ESRI Living Atlas. Raster tiles were mosaicked into a continuous surface for both study regions. Land cover was classified into eight categories, namely water, trees, flooded vegetation, crops, urban/built, bare ground, snow/ice, and rangeland (Figure 5).

To evaluate how buffer size influences land-cover relationships, land-cover proportions were computed for three nested delineations, namely the full watershed above each site, the portion of that watershed falling within 50 kilometers of the sampling location, and

the portion falling within 5 kilometers. All three delineations were carried forward into the correlation analysis below.

Watershed Delineation

Methods

Purpose

Mapping land cover within specified sampled watersheds aids in determining the relative importance, if any, of land cover types and proximity to sampling locations to the bacterial communities in the river at sampling sites.

Data Acquisition and Preparation

Land cover data were obtained from the 2019 ESRI Sentinel-2 Global Land Cover dataset, accessed via the ESRI Living Atlas Land Cover Explorer. Land cover rasters were mosaicked to create a continuous dataset for the study area.

Digital elevation model (DEM) data for watershed delineation were sourced from the HydroSHEDS database (Lehner et al., 2008). DEM data were imported into ArcGIS Pro for hydrological analysis. All spatial data were projected to appropriate coordinate systems: Mississippi River basin data were projected to *WGS_1984_UTM_Zone_15N*, and Yamuna River basin data were projected to *WGS 1984*.

Watershed Delineation

Watershed boundaries were delineated using standard hydrological tools in ArcGIS. First, the *Fill* tool was applied to the DEM to remove sinks and ensure continuous flow paths. Flow direction was computed from the filled DEM using the *Flow Direction* tool, followed by *Flow Accumulation* to identify drainage patterns. Pour points were set at designated sampling sites, and the *Watershed* tool was used to delineate catchments corresponding to each sampling location. All analyses employed the D8 flow direction algorithm.

Stream Network Extraction

Stream networks were derived from flow accumulation rasters using attribute-based extraction. Cells with flow accumulation values ≥ 8 were selected and converted to integer format. This semi-arbitrary value was set at 8 because that was the flow accumulation value of the Yamuna

headwaters sampling location nearest to the top of the watershed of the Yamuna River. The resulting raster was converted to polylines using the *Raster to Polyline* tool. A *Pairwise Buffer* of 1 km was applied to stream polylines, and buffers were clipped to watershed boundaries for targeted sampling zones.

Land Use Data Extraction

Our goal is to extract the land cover dataset that most accurately influences the bacterial communities in the rivers at each sampling location. There are many factors that contribute to the bacterial communities and we wanted to explore whether or not land cover was a major influencer. The connection between land and water varies between different watersheds and can vary significantly within the same watershed. We chose to explore this connection be defining the landcover spatial extent using six different methods, divided into two types, Type A and Type B.

Type A land cover extractions were mapped by including all the land within the watershed at three different extents: (1) The full watershed above the sampling location, (2) all watershed land within 50 km of the sampling site, and (3) all watershed land within 5 km of the sampling location.

Type B land cover extractions were mapped by only including land that was within 1km of watercourses. Here we define a watercourse as having a flow accumulation value of 8 or greater during our flow accumulation mapping process. Landcover data was extracted from the watercourse dataset with the same 3 extents used in the Type A extraction: (1) The full watershed above the sampling location, (2) all watershed land within 50 km of the sampling site, and (3) all watershed land within 5 km of the sampling location.

Correlation Analysis Between ASVs and Land Use

Pearson correlation coefficients were computed between the relative abundances of the 29 SIMPER-identified taxa and the eight land-cover classes using corr.test() in the *psych* package (Revelle 2025). Correlations were computed separately for each of the three delineations. To control for multiple testing, p-values were adjusted to control the false discovery rate using a Benjamini-Hochberg correction (Benjamini and Hochberg 1995) across the full set of correlations. Figures were produced with *ggplot2* (Wickham 2016).

Results:

Sampling Design

We compared surface-water bacterial communities from two major river systems, the Upper Yamuna in India and the Upper Mississippi at its headwaters in Lake Itasca, MN (Figure 1A), together with the Yamuna headwaters below the Yamunotri Glacier in Uttarakhand, India (Figure 1B, 1C).

Bacterial Community Clustering

Clustering reduced the eleven sites to a small number of community types, so that the taxa distinguishing one type from another could be identified in the next step. k-means clustering of the site-by-taxon relative abundance matrix identified five bacterial community types (Table 2). Cluster assignment closely matched geography, with each cluster representing a distinct region. The clusters were the Mississippi River, the Yamuna River at Sangam, Tons/Paonta Sahib, New Delhi, and the Yamuna River headwaters. The two Mississippi River sites, separated by hundreds of river kilometers, clustered together and apart from every Yamuna River site, while the two New Delhi sites, collected 6 meters apart, formed their own cluster as well.

Identifying the Set of Indicator Bacteria

To find which taxa distinguish one community cluster from another, we applied SIMPER, which ranks taxa by how much each contributes to the Bray-Curtis dissimilarity between clusters. Contributions to between-cluster dissimilarity appeared to level off at approximately 1% in all ten pairwise comparisons among the five clusters (Figure 2), so we used 1% as a threshold for identifying indicator taxa, meaning the taxa responsible for separating one cluster from another. Twenty-nine taxa exceeded that threshold (Table 3).

Geobacter and Arcobacter contributed most to the differences among clusters. Each accounted for approximately 7% of the average Bray-Curtis dissimilarity, and together the two genera account for roughly a seventh of all dissimilarity. Both were abundance at the two New Delhi sites and nearly absent everywhere else (Figure 3). The hgcI clade

(Actinobacteria), the family Burkholderiaceae, and the genus Flavobacterium were the next largest contributors to average dissimilarity (Table 3).

Each cluster carried a distinct fingerprint., meaning a distinct pattern of relative abundances across the 29 indicator taxa. New Delhi was dominated by Geobacter, Arcobacter, Tolumonas, Macellibacteroides fermentans, Arcobacter cryaerophilus, and Sulfurimonas. The Mississippi River cluster was dominated by the hgcI clade and SAR11 Clade III, and the Sangam by the hgcI clade and the CL500-29 marine group. Tons/Paonta Sahib was dominated by Flavobacterium, Burkholderiaceae, and Pseudarcicella, and the Yamuna River headwaters by Flavobacterium, Chloroplast reads, Polaromonas, Methylotenera, Rhodoferax, and Ferruginibacter (Figure 3).

Correlations Between Indicator Taxa and Land Cover

We computed Pearson correlations between the relative abundances of the 29 indicator taxa and the eight land-cover classes, separately for the 5-kilometer buffer, the 50-kilometer buffer, and the whole upstream watershed. Each correlation paired one taxon with one land-cover class. P-values were adjusted to control the false discovery rate at 0.05 using a Benjamini-Hochberg correction across the full set of tests, and we refer to any taxon and land-cover pair that remained significant as an association. Results are displayed as a heatmap comparing the three delineations (Figure 5)., with taxa ordered by average contribution to Bray-Curtis dissimilarity and cell color indicating the direction and significance of each correlation. Buffer size strongly influenced the strength of the correlations. Within the 5-kilometer buffer, correlations were weak and no associations remained significant after correction. At the 50-kilometer buffer, twenty associations were significant, all of them positive. Over the whole upstream watershed, eighteen were significant with sixteen being positive and two negatives.

The significant associations fell into four indicator sets, each defined by the land-cover class(es) its taxa correlated with (Table 4). An urban set of Geobacter, Arcobacter, Arcobacter cryaerophilus, Sulfurimonas, Tolumonas, and Macellibacteroides fermentans correlated positively with urban and built cover. Three of the six were significant at both the 50-kilometer buffer and whole upstream watershed, and the other

three at the 50-kilometer buffer alone. A glacial headwater set up of Polaromonas, Rhodoferax, Methylotenera, Ferruginibacter, and Chloroplast reads correlated positively with snow and ice cover. An open-water set of the hgcI clade, SAR11 Clade III, Sporichthyaceae, and Verrucomicrobiae correlated positively with water cover. The hgcI clade also correlated negatively with rangeland over the whole upstream watershed. Finally, a vegetation set correlated with the plant cover of the contributing land. Burkholderiaceae and Flavobacterium correlated positively with rangeland at 50-kilometer buffer. Limnohabitans and Sediminibacterium correlated positively with tree cover at both the 50-kilometer buffer and the whole upstream watershed, and Algoriphagus with tree cover over the whole upstream watershed. Dinghuibacter correlated negatively with rangeland over the whole upstream watershed.

Four further associations involved three taxa that we did not assign to a set. The CL500-29 marine group, OPB56, and the MWH-UniP1 aquatic group each correlated positively with crop cover over the whole upstream watershed, and the MWH-UniP1 aquatic group also with bare ground at 50-kilometer buffer. Five of the 29 taxa, namely Pseudarcicella, Hydrogenophaga, Fluviicola, the NS11-12 marine group, and Cyanobium PCC-6307, showed no significant association with any land-cover class at any delineation.

Discussion:

This study extends our earlier comparison of the Yamuna River and Mississippi River (Martinez et al. 2024. We set out to identify a small set of indicator taxa whose abundances vary with land cover in the contributing watershed, and we expected urban-influenced watersheds to carry higher abundances of pollutant-tolerant taxa. Clustering identified five community types, each corresponding to a distinct region, namely New Delhi, the Yamuna River headwaters, the Mississippi River, the Sangam, and Tons/Paonta Sahib. The position of a site along the river did not, by itself, determine cluster membership. The two Mississippi River sites clustered together despite hundreds of river kilometers between them, while the two New Delhi sites, 6 meters apart, formed a cluster distinct from every other Yamuna River site. The four mountain sites on the upper Yamuna River divided similarly. The Yamuna River

headwaters and Yamunotri town, 4 kilometers apart, formed one cluster, and the Tons River and Paonta Sahib sites, 24 kilometers apart, formed another.

Taxa as Indicators of Watershed Conditions

Geobacter and Arcobacter were the two largest contributors to between-cluster dissimilarity, each contributing roughly 7%, and were abundant only at the New Delhi sites. Lin et al. (2005) found that in an aquifer polluted by landfill leachate, pollution selected for members of the Geobacteraceae out of a much larger pool, and that the family made up about a quarter of the microbial community where iron reduction was the dominant redox process against 0.5% abundance in clean groundwater. Arcobacter was among the most abundant genera in the influent at fourteen municipal wastewater treatment plants and reached up to 30% of all bacteria in the treated effluent (Kristensen et al. 2020). The enrichment of both genera at New Delhi is consistent with conditions in that part of the river, where roughly 85% of the pollution entering the river is of domestic origin and dissolved oxygen at the Nizamuddin outfall can fall to zero (Sharma et al. 2024). Where dissolved oxygen reaches zero, aerobic respiration ceases, and the organisms that remain are those able to respire other electron acceptors or ferment.

Four further taxa at the New Delhi sites, namely Sulfurimonas, Macellibacteroides fermentans, Tolumonas, and Arcobacter cryaerophilus, have described metabolisms that fit the same conditions. Sulfurimonas oxidizes reduced sulfur compounds including sulfite, and members of the genus are recovered from redox clines, the boundary between oxygenated and sulfidic water (Han and Perner 2015). Macellibacteroides fermentans was isolated from an anaerobic filter treating abattoir wastewater and is an obligate anaerobe that grows only by fermentation, using none of the usual alternatives to oxygen such as nitrate or sulfate as an electron acceptor (Jabari et al. 2012). Tolumonas auensis was isolated from anoxic freshwater lake sediment and grows there by fermentation, although the organism also grows in the presence of oxygen (Fischer-Romero et al. 1996). Arcobacter cryaerophilus was the most abundant Arcobacter species in both influent and treated effluent in survey by Kristensen et al. (2020). Thus, these taxa enriched at New Delhi are not six unrelated microbes but instead are six organisms jointly characteristic of wastewater and oxygen-poor environments.

Members of the Burkholderiaceae also contributed strongly to between-cluster dissimilarity and were positively associated with rangeland at the 50-kilometer buffer. Under SILVA v138, Comamonadaceae are merged into the Burkholderiaceae, so Limnohabitans, Hydrogenophaga, Polaromonas, and Rhodoferax fall into the same family (Table 3). We therefore interpret the Burkholderiaceae at the order Burkholderiales. Burkholderiales carry an unusual capacity to break down aromatic compounds. Pérez-Pantoja et al. (2012) examined 80 Burkholderiales genomes and found nearly all central ring-cleavage pathways described in bacteria, together with most of the peripheral pathways feeding into them. Lignin, the ring-structured polymer that makes up much of plant material, is one such compound.

Rangeland in the ESRI Sentinel-2 product (Karra et al. 2021) is grass and scrub cover, classified by what grows there rather than by how the land is managed, so rangeland supplies plant material whether or not it is grazed. Where rangeland is grazed, it also delivers sediment, nutrients, and faecal materials to streams (Agouridis et al. 2005), and manure application adds short-lived pulses of labile dissolved organic matter to field runoff in the one to three weeks that follow (Singh et al. 2014). High flow events are the likely route into the river for these nutrients, faecal matter, and organisms, all influencing the composition of the downstream watershed. Caillon et al. (2021) found that roughly 70% of the bacterial operational taxonomic units in streams and stream biofilms were shared in soil water, the water held in the spaces between soil particles, and that bacterial diversity in headwater streams was higher during high flow than during low flow.

As an indicator, the Burkholderiaceae association is weaker than that of the urban set. The taxon is classified only to family, so the organisms behind the association cannot be named. It is also not restricted to a single cluster, and plant material enters the water through a variety of means. Newton et al. (2011) report that Burkholderiales, commonly regarded as a freshwater bacterium, also turn up in terrestrial and estuarine settings, reflecting the broad habitat range of the order rather than a preference for any one setting.

Flavobacterium showed the same positive association with rangeland. Cytophaga-Flavobacteria break down biopolymers such as cellulose and chitin, part of the high-molecular-weight fraction of dissolved organic matter, and they are enriched on particles of organic detritus (Kirchman 2002). Crump et al. (1999) found the same enrichment in the Columbia River estuary, where most Cytophaga sequences came from the particle-attached fraction. The trapped particles there are organic-rich and supply most of the food for the estuarine food web.

Grazed rangeland is one source of such particles. Grazing delivers mineral sediment, faecal materials, and nutrients to streams (Agouridis et al. 2005), and it is the faecal and other organic material, rather than the mineral sediment, that these bacteria can use. Kirchman (2002) reports that Cytophaga-Flavobacteria increased six to eightfold in a glacial stream after a storm washed organic materials in from the surrounding land. Chavarria et al. (2021) found Flavobacterium and Sediminibacterium elevated in cattle-pasture and silvopasture streams in lowland Panama, where cattle along the banks raised turbidity and suspended solids.

The hgcI clade was positively associated with water cover at the 50-kilometer buffer and negatively associated with rangeland over the whole upstream watershed. The negative association with rangeland is the opposite of what Burkholderiaceae and Flavobacterium showed. Warnecke et al. (2005) surveyed ten mountain lakes and found that a single actinobacterial lineage, acI, which corresponds to the hgcI clade under the SILVA naming conventions, made up more than 90% of the Actinobacteria and up to 70% of total bacterial abundance. Newton et al. (2011) report that increased nutrient concentrations generally select against freshwater Actinobacteria, and rangeland delivers nutrients to streams whether through grazing or manure application. A negative association between the hgcI clade and rangeland is consistent with that pattern.

The glacial headwater set reflects no anthropogenic land use. Polaromonas, Rhodoferax, Methylotenera, Ferruginibacter, and Chloroplast reads were all positively correlated with snow and ice cover, and all five were most abundant at the Yamunotri sites. Polaromonas is among the dominant bacteria of glacial ice and sediments worldwide

and is common in snow sampled at high altitude (Darcy et al. 2011). Rhodoferax ferrireducens is psychrotolerant, growing between 4 and 30 degrees Celsius (Finneran et al. 2003). Methylotenera mobilis grows only on methylamine, a one-carbon compound, using it as its sole source of energy, carbon, and nitrogen (Kalyuzhnaya et al. 2006). Snow and ice cover is almost entirely confined to the Yamunotri watershed, so the correlations in this set are confounded with geography. What this set does show however is that the analysis identified more than one pattern. The New Delhi taxa correlated with urban and built cover while the glacial headwater taxa correlated with snow and ice, so the results are not a single gradient running from impacted to unimpacted sites.

The Same Taxa Recur Across Independent Studies

The 29 taxa identified here recur in independent studies of other river systems. Fontaine et al. (2023) identified six top-ranked genus-level bioindicators along the Danube River, namely the hgcI clade, the CL500-29 marine group, Flavobacterium, Limnohabitans, Candidatus Methylopumilus, and Sediminibacterium. These six were selected as indicators of biological status, assessed with the macroinvertebrate-based Saprobic Index, and of chlorophyll a as a measure of eutrophication. Five of the six are among the 29 taxa identified here, and three of those five rank within our eight largest contributors to between-cluster dissimilarity. Extending the comparison to the full ranked set of fifteen genus-level candidates reported by Fontaine et al., nine are among our 29. She et al. (2024) likewise report Limnohabitans, Flavobacterium, Pseudarcicella, and the hgcI clade as the most abundant riverine genera in urban and agricultural watersheds in China. Chavarria et al. (2021) recover Flavobacterium, Sediminibacterium, Pseudarcicella, and Hydrogenophaga as indicator taxa separating pasture from forest streams in lowland Panama. Ahlinder et al. (2024) link the hgcI clade, the CL500-29 marine clade, Limnohabitans, Sediminibacterium, and Rhodoferax to the inflows of five Swedish lakes. Four studies across three continents, using different sampling designs and methods, arrive at overlapping sets of taxa.

Buffer Size Matters

The strength of the taxon and land-cover correlations depended on the size of the delineation. No correlation survived correction within 5 kilometers of the sampling site,

twenty were significant at the 50-kilometer buffer, and eighteen over the whole upstream watershed. The three delineations are nested, so the whole watershed contains the 50-kilometer buffer, which in turn contains the 5-kilometer buffer. The comparison therefore does not separate near land from distant land. What it does show though is that land cover summarized over a larger area corresponded more closely to taxon abundance than land cover summarized over a smaller one.

Other work reports the same direction. Ahlinder et al. (2024) found that the presence of buildings, farms, and fields was associated with roughly a third of the most common taxa in lake inflows, and that taxa from those inflows were then detected in the lakes themselves. Weigel et al. (2023) found that agricultural and forested land use in boreal Finnish streams appeared most strongly in conductivity, pH, and phosphorus concentration, and that those local conditions in turn shaped which taxa were present, placing the origin of the effect upstream of where the sample was taken.

Runoff is generated unevenly across a watershed because wet headwaters often sit above drier developed lowlands. A watershed with 20% urban land therefore does not necessarily generate 20% of its runoff from that urban land (Turner et al. 2021). Land cover measured as a share of area, as it is here, is not the same then as land cover weighted by where the runoff came from.

Limitations

Several limitations temper these conclusions. This is a pilot study of eleven sites. Eleven observations are enough to detect the strong, consistent associations reported here, but not enough to fit or validate a predictive model of community composition on land cover. With eight land-cover classes and eleven sites, such a model has almost no residual degrees of freedom, and holding sites out to estimate its performance is not an option. Additionally, correlations estimated on eleven observations also have wide confidence intervals even where they have survived correction, so individual coefficients should not be overread. Further, samples come from a single visit in July and August of 2019, so land-cover effects cannot be separated from seasonal ones. Season is known to structure bacterial communities in the Upper Mississippi River

(Staley et al. 2015). The sequencing material also came from surface water at a single depth, so benthic sediment and biofilms were not sampled, and these can differ substantially from the planktonic fraction (Lennert et al. 2024; Staley et al. 2016). The particle load in surface water also varied among sites, so the proportion of particle-associated organisms captured was not consistent across the eleven samples.

Future Directions

Three extensions follow from these limitations. Land cover could be weighted by how water travels to the sampling point, using flow distance or travel time, rather than the unweighted buffers and watershed used here. Spatial stream network models provide a readily available framework for a weighted approach (Hsu et al. 2023), and the comparison would show whether a weighted delineation describes taxon abundance better than an unweighted one. Sampling repeatedly across seasons and flow states would separate land cover from seasonal and hydrological variation. Capturing seasons of high flow would be of value because that is when terrestrial bacteria enter the river in the largest numbers (Cailon et al. 2021). More sites would also make community-level modeling possible, and it would also allow us to test the indicator set proposed here on sites that were not used to identify it.

Conclusion

Bacterial communities at eleven sites on the Yamuna River and the Mississippi River fall into five clusters, each corresponding to a region. Sites on the Yamuna River appear in four of those five types, so being on the same river did not, by itself, make a community similar. Twenty-nine taxa account for most of the dissimilarity among the five clusters, and their abundances correlate with land cover in four sets. Geobacter and Arcobacter correlate with urban and built cover. Polaromonas and Rhodoferax correlate with snow and ice. The hgcI clade and SAR11 Clade III correlate with open water. Flavobacterium and Burkholderiaceae correlate with rangeland and tree cover. Many of these taxa appear in independent studies of elsewhere. The hgcI clade, Flavobacterium, Limnohabitans, and Sediminibacterium are among the bioindicators reported for the Danube River, for Chinese urban and agricultural watersheds, for Panamanian pasture streams, and for Swedish lake inflows.

The correlations depended on the size of the delineation over which land cover was summarized. Within 5 kilometers of the sampling site, none remained significant after multiple-testing correction, while twenty did at the 50-kilometer bugger and eighteen over the whole upstream watershed. These findings suggest then that riverine bacterial communities are most useful as indicators of conditions across the contributing watershed rather than of conditions at the bank.

Data Availability

Raw sequence data are available from the NCBI Sequence Read Archive under BioProject accession PRJNA1077596. Land-cover data are from the 2019 ESRI Sentinel-2 Global Land Cover product (Karra et al. 2021); elevation data are from HydroSHEDS (Lehner et al. 2008). Analysis code and derived data tables are available from corresponding author via request.

References

Carmen T. Agouridis, Stephen R. Workman, Richard C. Warner, and Gregory D. Jennings. Livestock grazing management impacts on stream water quality: A review. JAWRA Journal of the American Water Resources Association, 41(3):591-606, 2005. doi: 10.1111/j.1752-1688.2005.tb03757.x.

Jon Ahlinder, Karolina Ida Anna Eriksson, Maria Hellmér, Emelie Salomonsson, Malin Granberg, Ingrid Dacklin, Josefine Elving, and Björn Brindefalk. Upstream land use with microbial downstream consequences: Iron and humic substances link to legionella spp. Water Research, 256:121579, 2024. doi: 10.1016/j.watres.2024.121579.

Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. Journal of the Royal Statistical Society: Series B (Methodological), 57(1):289-300, 1995. doi: 10.1111/j.2517-6161.1995.tb02031.x.

Florian Caillon, Katharina Besemer, Peter Peduzzi, and Jakob Schelker. Soil microbial inoculation during flood events shapes headwater stream microbial communities and diversity. Microbial Ecology, 82(3):591-601, 2021. doi: 10.1007/s00248-021-01700-3

Benjamin J. Callahan, Paul J. McMurdie, Michael J. Rosen, Andrew W. Han, Amy Jo A. Johnson, and Susan P. Holmes. DADA2: High-resolution sample inference from Illumina amplicon data. Nature Methods, 13(7):581-583, 2016. doi: 10.1038/nmeth.3869.

Karina A. Chavarria, Kristin Saltonstall, Jorge Vinda, Jorge Batista, Megan Lindmark, Robert F. Stallard, and Jefferson S. Hall. Land use influences stream bacterial communities in lowland tropical watersheds. Scientific Reports, 11(1):21752, 2021. doi: 10.1038/s41598-021-01193-7.

K. Robert Clarke. Non-parametric multivariate analyses of changes in community structure. Australian Journal of Ecology, 18(1):117-143, 1993. doi: 10.1111/j.1442-9993.1993.tb00438.x.

Stéphane Compant, Jerzy Nowak, Tom Coenye, Christophe Clément, and Essaïd Ait Barka. Diversity and occurrence of Burkholderia spp. in the natural environment. FEMS Microbiology Reviews, 32(4):607-626, 2008. doi: 10.1111/j.1574-6976.2008.00113.x.

James B. Cotner and Bopaiah A. Biddanda. Small players, large role: Microbial influence on biogeochemical processes in pelagic aquatic ecosystems. Ecosystems, 5(2):105-121, 2002. doi: 10.1007/s10021-001-0059-3.

Byron C. Crump, E. Virginia Armbrust, and John A. Baross. Phylogenetic analysis of particle-attached and free-living bacterial communities in the Columbia River, its estuary, and the adjacent coastal ocean. Applied and Environmental Microbiology, 65(7):3192-3204, 1999. doi: 10.1128/AEM.65.7.3192-3204.1999

Chansotheary Dang, Elliott Kellner, Gregory Martin, Zachary B. Freedman, Jason Hubbart, Kirsten Stephan, Charlene N. Kelly, and Ember M. Morrissey. Land use intensification destabilizes stream microbial biodiversity and decreases metabolic efficiency. Science of The Total Environment, 767:145440, 2021. doi: 10.1016/j.scitotenv.2021.145440.

Xiongde Dong, Leyun Yang, Laura Sofie Harbo, Xinyu Yan, Ji Chen, Cancan Zhao, Yutong Xiao, Hao Liu, Shilin Wang, Yuan Miao, Dong Wang, and Shijie Han. Effects of land use on soil microbial community structure and diversity in the Yellow River floodplain. Journal of Plant Ecology, 16(1):rtac075, 2023. doi: 10.1093/jpe/rtac075.

Esri Inc. ArcGIS Pro. Environmental Systems Research Institute, Redlands, California, 2024. URL https://www.esri.com/en-us/arcgis/products/arcgis-pro/. Version 3.x.

Laurent Fontaine, Lorenzo Pin, Domenico Savio, Nikolai Friberg, Alexander K. T. Kirschner, Andreas H. Farnleitner, and Alexander Eiler. Bacterial bioindicators enable biological status classification along the continental Danube river. Communications Biology, 6:862, 2023. doi: 10.1038/s42003-023-05237-8.

John A. Hartigan and Manchek A. Wong. Algorithm AS 136: A k-means clustering algorithm. Journal of the Royal Statistical Society: Series C (Applied Statistics), 28(1):100-108, 1979. doi: 10.2307/2346830.

Syrie M. Hermans, Hannah L. Buckley, Bradley S. Case, Fiona Curran-Cournane, Matthew Taylor, and Gavin Lear. Using soil bacterial communities to predict physico-chemical variables and soil quality. Microbiome, 8:79, 2020. doi: 10.1186/s40168-020-00858-1.

Syrie Hermans, Anju Gautam, Gillian D. Lewis, Martin Neale, Hannah L. Buckley, Bradley S. Case, and Gavin Lear. Exploring freshwater stream bacterial communities as indicators of land use intensity. Environmental Microbiome, 19:45, 2024. doi: 10.1186/s40793-024-00588-z.

Stuart Findlay. Stream microbial ecology. Journal of the North American Benthological Society, 29(1):170-181, 2010. doi: 10.1899/09-023.1.

Tom J. Battin, Katharina Besemer, Mia M. Bengtsson, Anna M. Romani, and Aaron I. Packmann. The ecology and biogeochemistry of stream biofilms. Nature Reviews Microbiology, 14(4):251-263, 2016. doi: 10.1038/nrmicro.2016.15.

Tom J. Battin, Ronny Lauerwald, Emily S. Bernhardt, Enrico Bertuzzo, Lluís Gómez Gener, Robert O. Hall Jr, Erin R. Hotchkiss, Taylor Maavara, Tamlin M. Pavelsky, Lishan

Ran, Peter Raymond, Judith A. Rosentreter, and Pierre Regnier. River ecosystem metabolism and carbon biogeochemistry in a changing world. Nature, 613(7944):449-459, 2023. doi: 10.1038/s41586-022-05500-8.

Robert H. Hilderbrand, Ted Bambakidis, and Byron C. Crump. The roles of microbes in stream restorations. Microbial Ecology, 85(3):853-861, 2023. doi: 10.1007/s00248-023-02179-w.

Tsung-Ta David Hsu, Danlin Yu, and Meiyin Wu. Predicting fecal indicator bacteria using spatial stream network models in a mixed-land-use suburban watershed in New Jersey, USA. International Journal of Environmental Research and Public Health, 20(6):4743, 2023. Doi: 10.3390/ijerph20064743.

Krishna Karra, Caitlin Kontgis, Zoe Statman-Weil, Joseph C. Mazzariello, Mark Mathis, and Simon P. Brumby. Global land use / land cover with Sentinel-2 and deep learning. In 2021 IEEE International Geoscience and Remote Sensing Symposium (IGARSS), pages 4704-4707, 2021. doi: 10.1109/IGARSS47720.2021.9553499.

David L. Kirchman. The ecology of cytophaga-flavobacteria in aquatic environments. FEMS Microbiology Ecology, 39(2):91-100, 2002. doi: 10.1111/j.1574-6941.2002.tb00910.x.

Jannie Munk Kristensen, Marta Nierychlo, Mads Albertsen, and Per Halkjær Nielsen. Bacteria from the genus arcobacter are abundant in effluent from wastewater treatment plants. Applied and Environmental Microbiology, 86(9):e03044-19, 2020. doi: 10.1128/AEM.03044-19

Pierre Legendre and Eugene D. Gallagher. Ecologically meaningful transformations for ordination of species data. Oecologia, 129(2):271-280, 2001. doi: 10.1007/s004420100716.

Bernhard Lehner, Kristine Verdin, and Andy Jarvis. New global hydrography derived from spaceborne elevation data. Eos, Transactions American Geophysical Union, 89(10):93-94, 2008. Doi: 10.1029/2008EO100001.

Kinga J. Lennert, Andrea K. Borsodi, Dóra Anda, Gergely Krett, Péter B. Kós, and Attila I. Engloner. The effect of urbanization on planktonic and biofilm bacterial communities in different water bodies of the Danube River in Hungary. Scientific Reports, 14:23881, 2024. doi: 10.1038/s41598-024-75863-7.

Bin Lin, Martin Braster, Boris M. van Breukelen, Henk W. van Verseveld, Hans V. Westerhoff, and Wilfred F. M. Röling. Geobacteraceae community composition is related to hydrochemistry and biodegradation in an iron-reducing aquifer polluted by a neighboring landfill. Applied and Environmental Microbiology, 71(10):5983-5991, 2005. doi: 10.1128/AEM.71.10.5983-5991.2005.

Fangzhou Ma, Chenbin Wang, Yanjing Zhang, Jing Chen, Rui Xie, and Zhanbin Sun. Development of microbial indicators in ecological systems. International Journal of Environmental Research and Public Health, 19(21):13888, 2022. doi: 10.3390/ijerph192113888.

Marcel Martin. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, 17(1):10-12, 2011. doi: 10.14806/ej.17.1.200.

Osvaldo Martinez, Silas R. Bergen, and Jacob B. Gareis. Comparison of Yamuna (India) and Mississippi River (United States of America) bacterial communities reveals greater diversity below the Yamunotri Glacier. PLOS ONE, 19(7):e0304664, 2024. doi: 10.1371/journal.pone.0304664.

Raj Morya, Madan Kumar, Shashi Shekhar Singh, and Indu Shekhar Thakur. Genomic analysis of Burkholderia sp. ISTR5 for biofunneling of lignin-derived compounds. Biotechnology for Biofuels, 12:277, 2019. doi: 10.1186/s13068-019-1606-5.

Jari Oksanen, Gavin L. Simpson, F. Guillaume Blanchet, Roeland Kindt, Pierre Legendre, Peter R. Minchin, R. B. O'Hara, Peter Solymos, M. Henry H. Stevens, Eduard Szoecs, Helene Wagner, et al. vegan: Community Ecology Package, 2025. URL https://CRAN.R-project.org/package=vegan. R package version 2.6-10.

Yuwei Pan, Jiawei Xie, Weixing Yan, Tian C. Zhang, and Chongjun Chen. Response of microbial community to different land-use types, nutrients and heavy metals in urban river sediment. Journal of Environmental Management, 321:115855, 2022. doi: 10.1016/j.jenvman.2022.115855.

Trishala K. Parmar, Deepak Rawtani, and Y. K. Agrawal. Bioindicators: The natural indicator of environmental pollution. Frontiers in Life Science, 9(2):110-118, 2016. Doi: 10.1080/21553769.2016.1162753.

Danilo Pérez-Pantoja, Raúl Donoso, Loreine Agulló, Macarena Córdova, Michael Seeger, Dietmar H. Pieper, and Bernardo González. Genomic analysis of the potential for aromatic compounds biodegradation in burkholderiales. Environmental Microbiology, 14(5):1091-1117, 2012. Doi: 10.1111/j.1462-2920.2011.02613.x.

Xiaodong Qu, Ze Ren, Haiping Zhang, Min Zhang, Yuhang Zhang, Xiaobo Liu, and Wenqi Peng. Influences of anthropogenic land use on microbial community structure and functional potentials of stream benthic biofilms. Scientific Reports, 7:15117. 2017. doi: 10.1038/s41598-017-15624-x.

Christian Quast, Elmar Pruesse, Pelin Yilmaz, Jan Gerken, Timmy Schweer, Pablo Yarza, Jörg Peplies, and Frank Oliver Glöckner. The SILVA ribosomal RNA gene database

project: Improved data processing and web-based tools. Nucleic Acids Research, 41(D1):D590-D596, 2013. doi: 10.1093/nar/gks1219.

R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria, 2025. URL https://www.R-project.org/. Version 4.4.3.

William Revelle. Psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois, 2025. URL https://CRAN.R-project.org/package=psych. R package version 2.5.3.

M. Sagova-Mareckova, J. Boenigk, A. Bouchez, K. Cermakova, T. Chonova, T. Cordier, U. Eisendle, T. Elersek, S. Fazi, T. Fleituch, L. Frühe, M. Gajdosova, N. Graupner, A. Haegerbaeumer, A.-M. Kelly, J. Kopecky, F. Leese, P. Nõges, S. Orlic, K. Panksep, J. Pawlowski, A. Petrusek, J.J. Piggott, J.C. Rusch, R. Salis, J. Schenk, K. Simek, A. Stovicek, D.A. Strand, M.I. Vasquez, T. Vrålstad, S. Zlatkovic, M. Zupancic, and T. Stoeck. Expanding ecological assessment by integrating microorganisms into routine freshwater biomonitoring. Water Research, 191:116767, 2021. doi: 10.1016/j.watres.2020.116767.

Silvia Secchi and Moira McDonald. The state of water quality strategies in the Mississippi River Basin: Is cooperative federalism working? Science of The Total Environment, 677:241-249, 2019. doi: 10.1016/j.scitotenv.2019.04.381.

Madhuben Sharma, Sameeksha Rawat, Dheeraj Kumar, Amit Awasthi, Abhijit Sarkar, Atul Sidola, Tanupriya Choudhury, and Ketan Kotecha. The state of the Yamuna River: A detailed review of water quality assessment across the entire course in India. Applied Water Science, 14:175, 2024. doi: 10.1007/s13201-024-02227-x.

Yuanyang She, Peng Wang, Jiawei Wen, Mingjun Ding, Hua Zhang, Minghua Nie, and Gaoxiang Huang. Riverine bacterial communities are more shaped by species sorting in

intensive urban and agricultural watersheds. Frontiers in Microbiology, 15:1463549, 2024. doi: 10.3389/fmicb.2024.1463549.

Shatrughan Singh, Sudarshan Dutta, and Shreeram Inamdar. Land application of poultry manure and its influence on spectrofluorometric characteristics of dissolved organic matter. Agriculture, Ecosystems & Environment, 193:25-36, 2014. Doi: 10.1016/j.agee.2014.04.019.

Christopher Staley, Trevor J. Gould, Ping Wang, Jane Phillips, James B. Cotner, and Michael J. Sadowsky. Species sorting and seasonal dynamics primarily shape bacterial communities in the Upper Mississippi River. Science of The Total Environment, 505:435-445, 2015. doi: 10.1016/j.scitotenv.2014.10.012.

Christopher Staley, Trevor J. Gould, Ping Wang, Jane Phillips, James B. Cotner, and Michael J. Sadowsky. Sediments and soils act as reservoirs for taxonomic and functional bacterial diversity in the Upper Mississippi River. Microbial Ecology, 71(4):814-824, 2016. doi: 10.1007/s00248-016-0729-5.

Sean W. D. Turner, Jennie S. Rice, Kristian D. Nelson, Chris R. Vernon, Ryan McManamay, Kerim Dickson, and Landon Marston. Comparison of potential drinking water source contamination across one hundred U.S. cities. Nature Communications, 12:7254, 2021. doi: 10.1038/s41467-021-27509-9.

Thea Van Rossum, Michael A. Peabody, Miguel I. Uyaguari-Diaz, Kirby I. Cronin, Michael Chan, Jared R. Slobodan, Matthew J. Nesbitt, Curtis A. Suttle, William W. L. Hsiao, Patrick K. C. Tang, Natalie A. Prystajecky, and Fiona S. L. Brinkman. Year-long metagenomic study of river microbiomes across land use and water quality. Frontiers in Microbiology, 6:1405, 2015. doi: 10.3389/fmicb.2015.01405.

Qiong Wang, George M. Garrity, James M. Tiedje, and James R. Cole. Naïve Bayesian Classifier for Rapid Assignment of rRNA sequences into the New Bacterial Taxonomy. Applied and Environmental Microbiology, 73(16):5261-5267, 2007. doi: 10.1128/AEM.00062-07.

Falk Warnecke, Ruben Sommaruga, Raju Sekar, Julia S. Hofer, and Jakob Pernthaler. Abundances, identity, and growth state of actinobacteria in mountain lakes of different UV transparency. Applied and Environmental Microbiology, 71(9):5551-5559, 2005. doi: 10.1128/AEM.71.9.5551-5559.2005.

Benjamin Weigel, Caio Graco-Roza, Jenni Hultman, Virpi Pajunen, Anette Teittinen, Maria Kuzmina, Evgeny V. Zakharov, Janne Soininen, and Otso Ovaskainen. Local eukaryotic and bacterial stream community assembly is shaped by regional land use effects. ISME Communications, 3:65, 2023. doi: 10.1038/s43705-023-00272-2.

Hadley Wickham. ggplot2: Elegant Graphics for Data Analysis. Springer-Verlag, New York, 2nd edition, 2016. doi: 10.1007/978-3-319-24277-4.

Zhenyu Zhang, Junhao Tang, Long Wang, Chongchong Zhu, Qian Xun, and Daniel Rosado. Amplified impacts of human activities: Non-linear responses of riverine microbial communities to distribution of land use. Journal of Environmental Management, 373:123937, 2025. doi: 10.1016/j.jenvman.2024.123937.

Figures and Tables Included

Figure 1

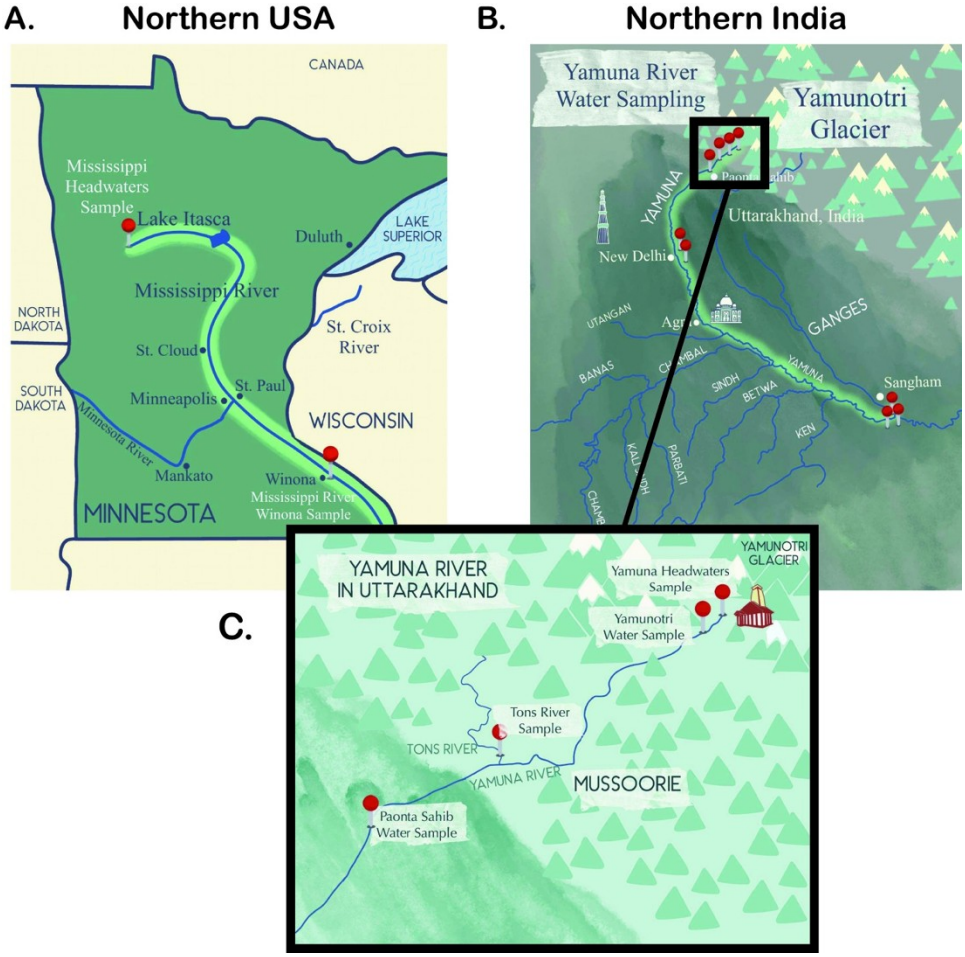


Table 1

Water sample source	Date water collected	Coordinates	Water sample temperature (Celsius)
Mississippi headwaters	7-13-19	47°14'22.8"N 95°12'28.0"W	27
Mississippi River at Winona	7-12-19	44°03'18.2"N 91°38'07.7"W	23.2
Yamuna headwaters	7-20-19	31°00'36.5"N 78°27'40.3"E	19.3
Yamuna River at Yamunotri	7-20-19	30°58'35.3"N 78°26'19.8"E	18.6
Tons River	7-21-19	30°30'38.0"N 77°49'03.1"E	25.6
Yamuna River at Paonta Sahib	7-21-19	30°26'04.1"N 77°37'14.9"E	25.3
Yamuna River in Delhi-1	7-22-19	28°40'33.8"N 77°13'54.7"E	32.9
Yamuna River in Delhi-2	7-22-19	28°40'32.3"N 77°13'55.4"E	32.9
Yamuna River upstream of Sangam	7-28-19	25°25'41.3"N 81°52'37.7"E	30.9
Ganga River upstream of Sangam	8-2-19	25°26'22.3"N 81°53'01.2"E	32.6
Yamuna and Ganga River at Sangam	7-28-19	25°25'36.1"N 81°53'13.8"E	31.1

Table 2

Cluster	Site
Mississippi	Mississippi Headwaters
	Mississippi River at Winona
Yamuna at Sangam	Ganga River at Sangam
	Yamuna River at Sangam
	Sangam River Mix
Tons/Paonta Sahib	Tons River near Yamuna River
	Yamuna River at Paonta Sahib
New Delhi	Yamuna River at New Delhi - 1
	Yamuna River at New Delhi - 2
Yamuna Headwaters	Yamunotri Town
	Yamuna Headwaters

Table 2. Site-Cluster Assignment

Figure 2

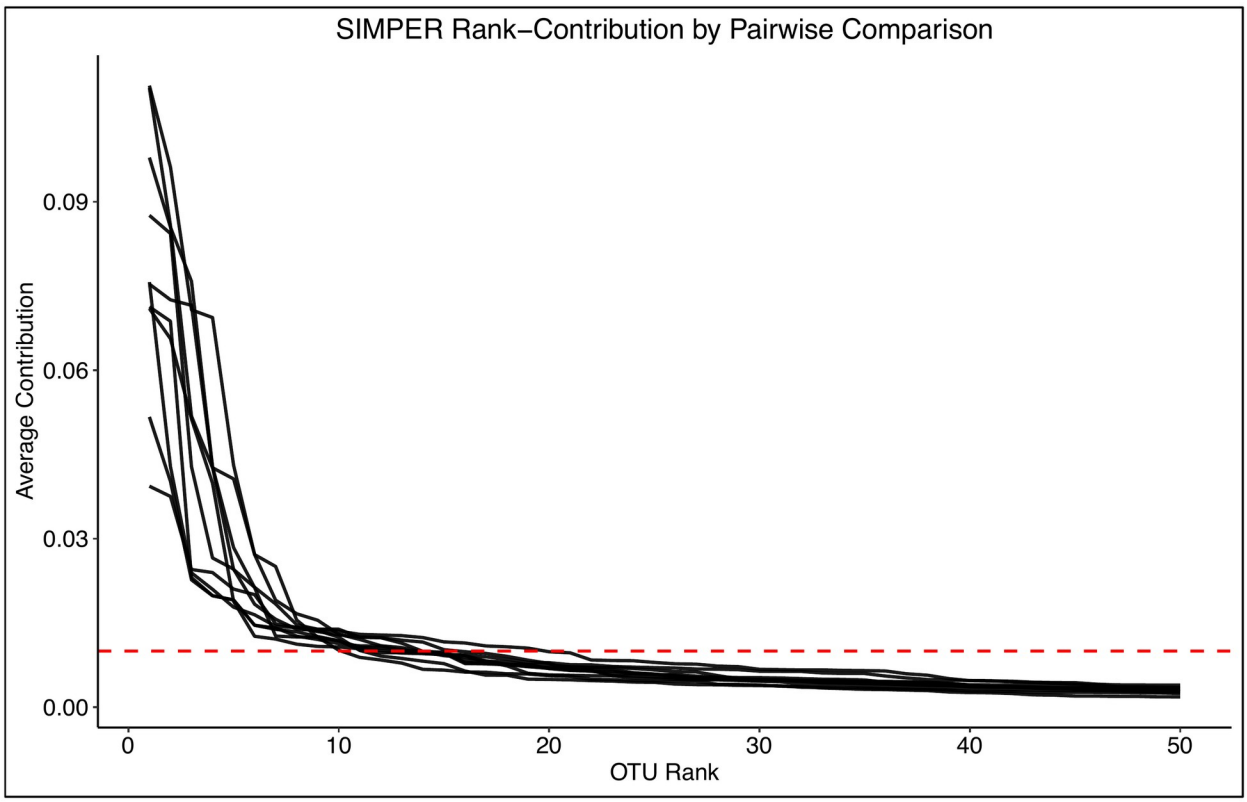


Table 3

Class;Order;Family;Genus;Species	Average Contribution
Deltaproteobacteria;Desulfuromonadales;Geobacteraceae;Geobacter;NA	0.0702
Campylobacteria;Campylobacterales;Arcobacteraceae;Arcobacter;NA	0.0701
Actinobacteria;Frankiales;Sporichthyaceae;hgcI_clade;NA	0.058
Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;NA;NA	0.055
Bacteroidia;Flavobacteriales;Flavobacteriaceae;Flavobacterium;NA	0.0542
Alphaproteobacteria;SAR11_clade;Clade_III;NA;NA	0.042
Acidimicrobiia;Microtrichales;Ilumatobacteraceae;CL500-29_marine_group;NA	0.0396
Bacteroidia;Cytophagales;Spirosomaceae;Pseudarcicella;NA	0.0286
Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Limnohabitans;NA	0.0234
Verrucomicrobiae;NA;NA;NA;NA	0.0212
Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Hydrogenophaga;NA	0.0188
Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Polaromonas;NA	0.0186
Bacteroidia;Flavobacteriales;Crocinitomicaceae;Fluviicola;NA	0.0176
Bacteroidia;Sphingobacteriales;NS11-12_marine_group;NA;NA	0.0166
Oxyphotobacteria;Chloroplast;NA;NA;NA	0.0146
Gammaproteobacteria;Betaproteobacteriales;Methylophilaceae;Methylotenera;NA	0.014
Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;Rhodoferax;NA	0.0139
Bacteroidia;Cytophagales;Cyclobacteriaceae;Algoriphagus;NA	0.0137
Bacteroidia;Chitinophagales;Chitinophagaceae;Sediminibacterium;NA	0.0135
Gammaproteobacteria;Aeromonadales;Aeromonadaceae;Tolumonas;NA	0.0129
Actinobacteria;Frankiales;Sporichthyaceae;NA;NA	0.0127
Ignavibacteria;OPB56;NA;NA;NA	0.0127
Bacteroidia;Bacteroidales;Tannerellaceae;Macellibacteroides;fermentans	0.0125
Oxyphotobacteria;Synechococcales;Cyanobiaceae;Cyanobium_PCC-6307;NA	0.012
Gammaproteobacteria;Betaproteobacteriales;Burkholderiaceae;MWH-UniP1_aquatic_group;NA	0.0109
Campylobacteria;Campylobacterales;Arcobacteraceae;Arcobacter;cryaerophilus	0.0108
Bacteroidia;Chitinophagales;Chitinophagaceae;Dinghuibacter;NA	0.0108
Campylobacteria;Campylobacterales;Thiovulaceae;Sulfurimonas;NA	0.0103
Bacteroidia;Chitinophagales;Chitinophagaceae;Ferruginibacter;NA	0.0102

Table 3. ASVs with ≥ 1% Average Contribution to Bray-Curtis Dissimilarity Score

Figure 3

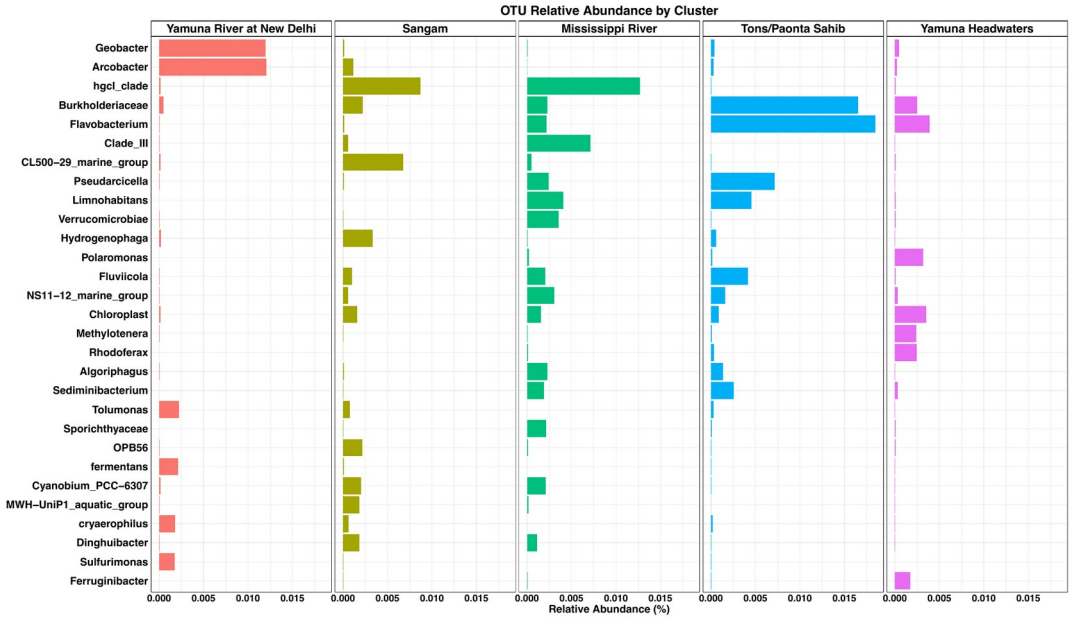


Figure 4

Figure 5

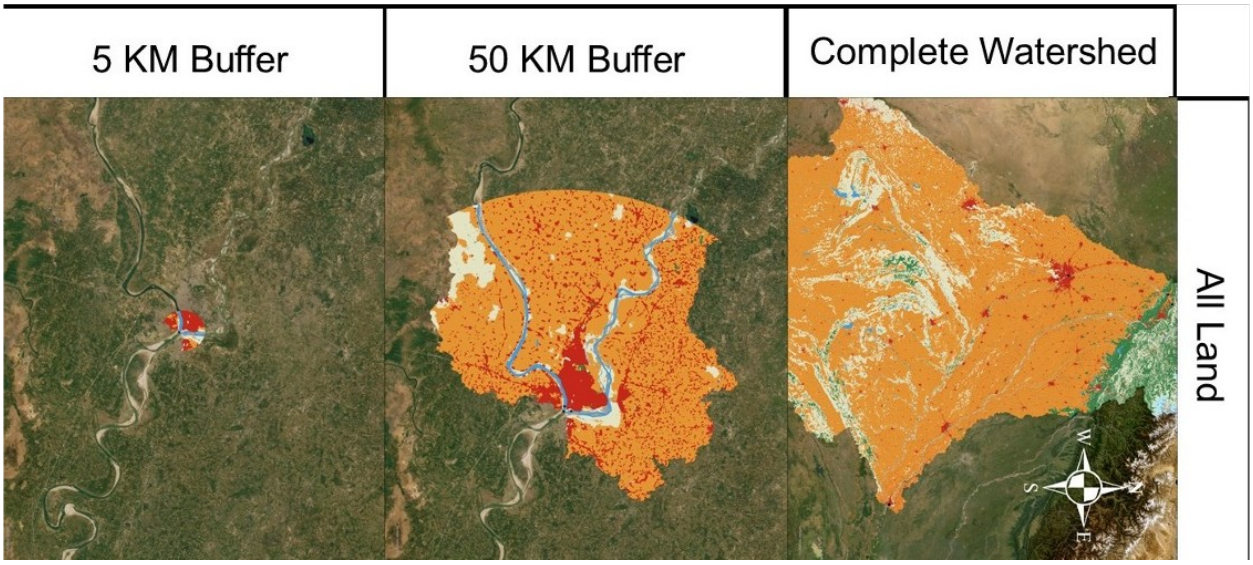
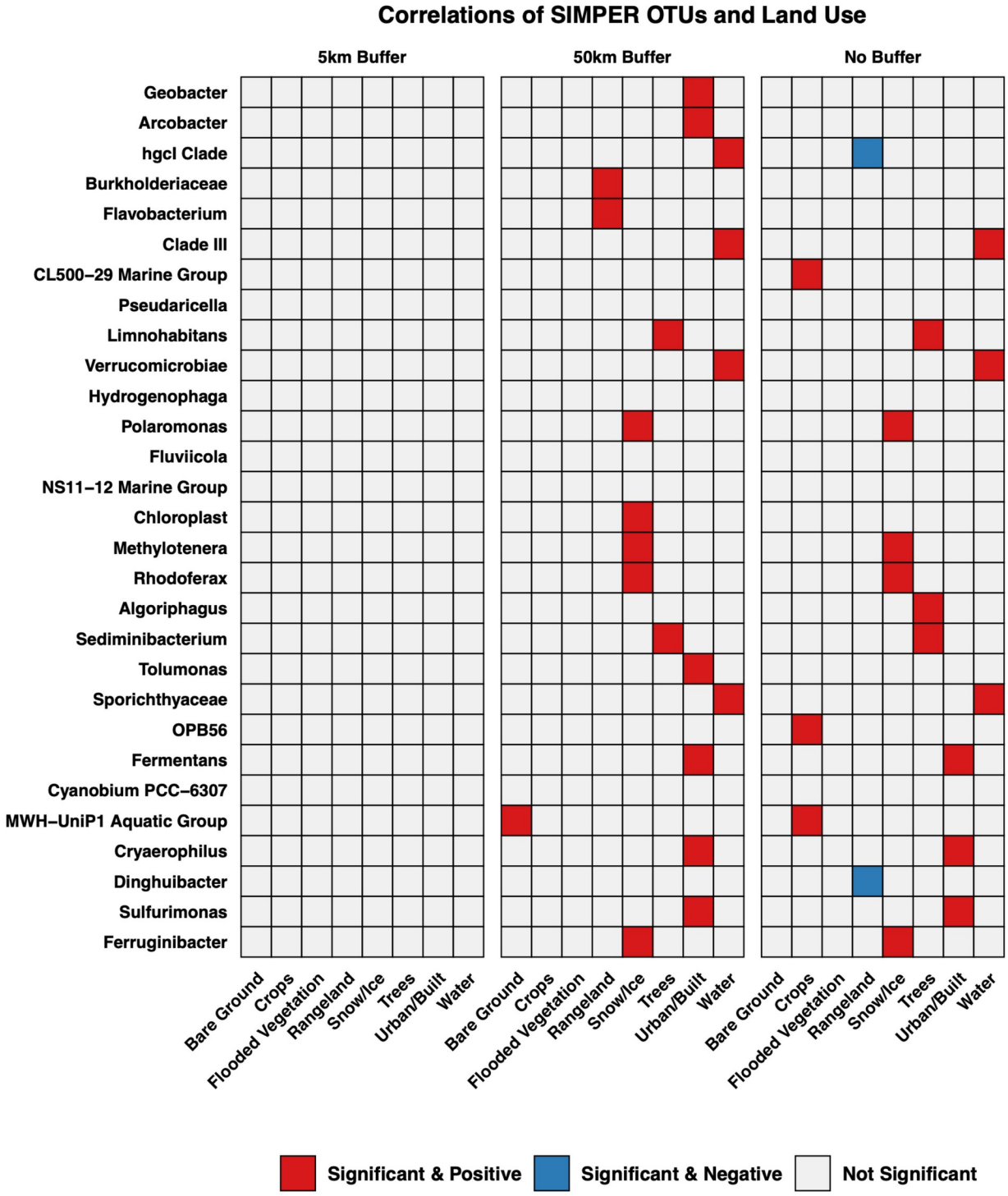


Table 4

Group	Taxon	Land Use Type	50 Kilometers	Whole Watershed
Urban Influence	Geobacter	Urban/Built	+	
	Arcobacter	Urban/Built	+	
	Cryaerophilus	Urban/Built	+	+
	Sulfurimonas	Urban/Built	+	+
	Tolumonas	Urban/Built	+	
	Fermentans	Urban/Built	+	+
Glacial Headwaters	Polaromonas	Snow/Ice	+	+
	Rhodoferax	Snow/Ice	+	+
	Methylotenera	Snow/Ice	+	+
	Ferruginibacter	Snow/Ice	+	+
	Chloroplast	Snow/Ice	+	
Open Water	hgcI Clade	Water	+	
	SAR11 Clade III	Water	+	+
	Sporichthyaceae	Water	+	+
	Verrucomicrobiae	Water	+	+
	hgcI Clade	Rangeland		-
Vegetation	Burkholderiaceae	Rangeland	+	
	Flavobacterium	Rangeland	+	
	Limnohabitans	Trees	+	+
	Sediminibacterium	Trees	+	+
	Algoriphagus	Trees		+
	Dinghuibacter	Rangeland		-
Other	MWH-UniP1 Aquatic Group	Bare Ground	+	
	MWH-UniP1 Aquatic Group	Crops		+
	CL500-29 Marine Group	Crops		+
	OPB56	Crops		+

Figure 6


Correlations of SIMPER OTUs and Land Use
5km Buffer
50km Buffer
No Buffer
Geobacter
Arcobacter
hgcI Clade
Burkholderiaceae
Flavobacterium
Clade III
CL500–29 Marine Group
Pseudaricella
Limnohabitans
Verrucomicrobiae
Hydrogenophaga
Polaromonas
Fluviicola
NS11–12 Marine Group
Chloroplast
Methylotenera
Rhodoferax
Algoriphagus
Sediminibacterium
Tolumonas
Sporichthyaceae
OPB56
Fermentans
Cyanobium PCC–6307
MWH–UniP1 Aquatic Group
Cryaerophilus
Dinghuibacter
Sulfurimonas
Ferruginibacter
Bare Ground
Crops
Flooded Vegetation
Rangeland
Snow/Ice
Trees
Urban/Built
Water
Significant & Positive
Significant & Negative
Not Significant