A functional microbiome catalog crowdsourced from North American rivers

Authors:
Mikayla A. Borton[1]*, Bridget B. McGivern[1], Kathryn R. Willi[2], Ben J. Woodcroft[3], Annika C. Mosier[4], Ted Bambakidis[5], Derick M. Singleton[4], Filipe Liu[6], Janaka N. Edirisinghe[6], José P. Faria[6], Ikaia Leleiwi[1], Rebecca A. Daly[1], Amy E. Goldman[7], Michael J. Wilkins[1], Ed K. Hall[2], Christa Pennacchio[8], Simon Roux[8,9], Emiley A. Eloe-Fadrosh[8,9], Matthew B. Sullivan[10], Christopher S. Henry[6], Elisha M. Wood-Charlson[9], Matthew R.V. Ross[2], Christopher S. Miller[4], Byron C. Crump[11], James C. Stegen[12,13], Kelly C. Wrighton[1]*

[1] Department of Soil and Crop Sciences, Colorado State University, Fort Collins, CO, USA
[2] Department of Ecosystem Science and Sustainability, Colorado State University, Fort Collins, CO, USA
[3] Centre for Microbiome Research, School of Biomedical Sciences, Queensland University of Technology, Translational Research Institute, Woolloongabba, Australia
[4] Department of Integrative Biology, University of Colorado Denver, Denver, CO, USA
[5] Department of Microbiology, Oregon State University, Corvallis, OR 97331, United States
[6] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA
[7] Earth Systems Science Division, Pacific Northwest National Laboratory, Richland, WA, USA
[8] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[9] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[10] Department of Microbiology, The Ohio State University, Columbus, OH, USA
[11] College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA
[12] Terrestrial and Aquatic Integration, Pacific Northwest National Laboratory, Richland, WA, USA
[13] School of the Environment, Washington State University, Pullman, WA, USA

Corresponding authors*: mborton@colostate.edu, wrighton@colostate.edu

Abstract: Predicting elemental cycles and maintaining water quality under increasing anthropogenic influence requires understanding the spatial drivers of river microbiomes. However, the unifying microbial determinants governing river biogeochemistry are hindered by a lack of genome-resolved functional insights and sampling across multiple rivers. Here we employed a community science effort to accelerate the sampling of river microbiomes to create the Genome Resolved Open Watersheds database (GROWdb). This resource profiled the identity, distribution, function, and expression of thousands of microbial genomes across rivers covering 90% of United States watersheds. We identified the most cosmopolitan microbiome members, while also revealing local drivers of strain endemism across ecological dimensions. We provide the first evidence that microbial functional trait expression followed the tenets of the River Continuum Concept, suggesting the structure and function of river microbiomes is predictable. GROWdb is a publicly available resource that paves the way for watershed predictive modeling and microbiome-based management practices.

Introduction

Rivers are a core element of the Earth system, acting as an important nexus of nutrient transport across terrestrial and aquatic interfaces. They are recognized hotspots for biogeochemical processes that impact global greenhouse gas (GHG) emissions, eutrophication, and acidification[1–4]. Processes occurring throughout rivers generate outcomes that are immediately relevant to society, including impacts on sustainable energy, agriculture, environmental health, and human health[5–7]. From local to global scales, the emergent functioning of rivers arises through the integration between hydrology and biogeochemistry. Functional attributes of river microbiomes remain largely unexplored, which undermines fundamental

understanding of their ecology and our ability to use a genomic foundation to predict river biogeochemical function.

There is a growing awareness that streams and rivers are significant contributors to global terrestrial carbon and nitrogen budgets[8,9], and that microbial metabolism controls much of these transformations. For instance, microbial metabolism in rivers has a major influence on carbon (C) respiration and sequestration, nitrogen (N) cycling and uptake, and greenhouse gas emissions[3,10–12]. Additionally, microbial biomass is a primary food supply for river food webs, and microbial processes that drive the cycling of biogeochemically active elements (e.g., C, N, P) and pollutants (e.g., Hg, As)[13,14]. Given these important contributions there is a growing need to better resolve the ecology, functional potential, and biogeochemical contributions of microbes across diverse river systems.

Despite being critical modulators of biogeochemistry, river microbiomes are largely under sampled in terms of the types of sequencing methods used and the distribution of sampling. For example, a majority of river microbiome studies rely only on 16S rRNA gene analysis[15–17]. While these single gene studies have advanced understanding of microbial community diversity and membership, they lack microbial functional information. Albeit less, there are studies with metagenomic data that provide functional attributes of river microbiomes, but rely heavily on databases rather than *de novo* assembly and metagenome assembled genomes[18], potentially masking the contributions of novel members of the microbiome. Even fewer studies employ genome-resolved expression methods. Additionally, in terms of distributed sampling, most studies to date focus on a single site or stream network[10,19], leaving the generalizability of the findings across river systems uncertain. To establish a transferable functional understanding of river microbiomes, there is a need to genomically resolve the taxonomy, metabolic potential, and expression of river microbiomes at scale.

To meet this need, we developed the Genome Resolved Open Watersheds database (GROWdb) effort, which aims to increase genomic sampling and understanding of global river microbiomes. An emphasis of GROWdb is to create a publicly available and ever-expanding microbial genome database that is focused on rivers while being interoperable with databases from other ecosystems. GROWdb is based on a network-of-networks approach[20] to move beyond a small collection of well-studied rivers, towards a spatially distributed, global network of systematic observations[21]. GROWdb represents the first microbial, river-focused resource parsed at various scales from genes to MAGs to community level including expression and potential based measurements that will be of interest to microbiologists, ecologists, geochemists, hydrologists, and modelers.

Results and Discussion

Construction of the Genome Resolved Open Watersheds database (GROWdb) enabled by community sampling and sequencing

Here we build and release the first version of the GROW database (GROWdb)[22], which focused on a crowd-sourced, distributed sampling of United States rivers, generating a total of 3.8Tbp of sequencing from these aquatic systems (**Fig. 1A**). To obtain an environmental context for each metagenome and metatranscriptome, we mapped sampling locations to watershed related databases, generating over 250 geospatial variables per site including watershed metrics, land use information, and climate data. Additionally, paired geochemical data were collected at each sampling location (**Fig. 1B**). We aimed to capture whole community-level variations in potential function and gene expression, placing these data into the context of geographic and environmental gradients across the United States.

The dataset for this study consists of metagenomic (n=158, 3.2 Tbp) and metatranscriptomic (n=57, 663 Gbp) samples collected at 163 distinct sampling locations across the United States (**Supplementary Data 1**). Surface water sampling covered 90% of United States watersheds (n=21 as determined by hydrologic unit 2) (**Fig. 1C**). Rivers sampled spanned a range of ecoregions, stream orders, and watershed sizes (**Supplementary Fig. 1**), enabling the fundamental ecological questions to be asked for the first time at a genome level. For example, we can utilize genome-resolved data with stream order to understand how microbial community substrate usage might change down a river network[16,23–25] or with land use metrics to decipher how human influence may impact microbial community function[7,26].

Collectively, GROWdb integrates genomics, biogeochemistry, and a broad range of contextual environmental variables into a platform that enable a predictive framework of microbiomes and their biogeochemical contributions.

Over 2,000 unique microbial genomes recovered from United States surface waters

To uncover the key microbial players and functions in surface water river microbiomes, we constructed a genome database composed of metagenome-assembled genomes (MAGs). Our sequencing represents on average of 3-fold more sequencing per sample compared to published riverine metagenome studies, thereby increasing the sensitivity for detecting the breadth of microbial functions encoded in these systems (**Supplementary Fig. 2**). From this sequencing, we assembled and reconstructed 3,825 medium and high-quality MAGs, which were dereplicated into 2,093 unique MAGs at 99% identity (**Fig. 2, Supplementary Data 2**). Based on read mapping, the majority (mean=52%) of metagenomic reads mapped back to this surface water derived MAG database, signifying that the underlying data were well represented by the database.

The dereplicated MAG database ($n = 2{,}093$) contained genomes from 27 phyla, many of which represent the most abundant and cosmopolitan lineages in rivers[16,17,27,28]. Using the Genome Taxonomy Database toolkit (GTDB-tk), we found that a subset of our genomes represented newly sampled lineages, including 10 families and 128 genera across 16 phyla (**Supplementary Fig. 2**). Additionally, a large proportion of our MAGs belonged to lineages defined only by alphanumeric identifiers (e.g., UBA) at the phylum (n=1), class (n=17), order (n=121), and family (n=196) levels (**Supplementary Fig. 2**). Highlighting the relevance of our database, analysis of 266,764 public metagenome datasets in the Sequence Read Archive (SRA)[29] revealed that species in GROWdb were detected in 90% of metagenomes classified as riverine and 46% of metagenomes classified as freshwater, aquatic, or riverine (based on matching taxonomical assignments). We verified the most prevalent phyla and genera in GROWdb had similar representation in publicly available metagenome set (**Supplementary Fig. 2**). Moreover, GROWdb species were detected from other environments including wastewater, lake water, sediment, marine, estuary, activated sludges, and soil, supporting the notion that rivers are truly integrators across landscapes (**Fig. 2B**).

The comparison to publicly available metagenomes underscored the need for this river-based microbiome studie, as freshwater related metagenomes were only half to a third as many as their soil and ocean counterparts respectively. Additionally, this analysis highlighted the importance of standardized metadata practices for data reuse, as more than 10% of metagenomes in the publicly available set had vague classifications such as ‘metagenome’ or ‘bacterium’, making the data nearly unusable. GROWdb ascribes to standardized protocols and metadata practices[30–32], making interoperability a hallmark of this resource and permitting meta-analysis with other studies, which is of utmost importance as our ability to scale multi-omics methods rapidly increases.

Aerobic and light-driven energy metabolisms are core and dominant across river microbiomes

We identified core and dominant features of riverine metagenomes and metatranscriptomes across rivers (**Fig. 3**). In terms of metagenome dynamics, members of the Actinobacteria, Proteobacteria, Bacteroidota, and Verrucomicrobiota dominated all samples (**Fig. 3A**). At lower taxonomic levels, genera that were the most cosmopolitan (occupancy; number of samples) across samples, were also the most dominant members of these communities (**Fig. 3B**). This was especially true for MAGs affiliated with the genus *Planktophilia*, a well-known freshwater microorganism by 16S rRNA gene analysis[33,34], that were present in 70% of the GROW metagenomes, and had the highest mean relative abundance at 12%. Five other genera including *Limnohabitans A*, *Polynucleobacter*, *Methylopumilus*, *Nanopelagicus*, and *Sediminibacterium* were also present in >50% of metagenomes.

For the subset of samples with paired metranscriptomes, we evaluated the microorganisms that were most transcriptionally active. To focus on the most relevant lineages, we limited our analyses to MAGs that were expressing genes in at least 10% of the samples. These resulted in a quarter of the 2,093 MAGs being considered active, including at least one representee from 19 of the 27 phyla displayed in

Fig. 2. The six most pertinent genera identified by metagenomics (**Fig. 3B**) also belonged to the top 25 genera with the highest mean gene expression (**Fig. 3C**), indicating prevalence, dominance, and activity were in agreement in these samples. Furthermore, three of these pertinent lineages (*Methylopumilus, Polynucleobacter, Planktophilia),* as well as members of Pirellula B, and two alphanumeric genera of Burkholderiaceae (UBA3064, UBA954) were transcriptionally active in every metatranscriptome, here denoted as the core, active genera. Notably, this was not an aggregate genus level effect, as each of these genera apart from *Polynucleobacter* had a single MAG representative that was expressed in every metatranscriptome. Our goal for these analyses was to constrain the thousands of genomes in GROWdb to a most-wanted list of the most active microorganisms in river systems, identifying lineages that could represent diagnostic or metabolism targets needing accurate representation in biogeochemical models moving forward.

To understand the impacts of these core active genera in modulating river geochemistry, we used genomic content to assign metabolic traits to each MAG, inventorying the capacity to use oxygen, light, nitrogen, sulfur and other key energy generation systems (**Supplementary Fig. 3, Supplementary Data 3**). We found that the core and most expressed genera had the capacity for aerobic respiration and the use of light as an energy source, capturing energy via high-yield photosystems (oxygenic or anoxygenic) or simple, low yield photorhodopsins. In fact, of the top 25 most active genera, more than 90% were capable of aerobic respiration and light driven metabolisms, with many encoding multiple light harvesting mechanisms (**Fig. 3C, Supplementary Fig. 4**). In addition to heterotrophy and autotrophy, many of these core active lineages had the capacity for aerobically oxidizing inorganic electron donors like sulfur and possibly methane, the latter via a divergent particulate methane monooxygenase (**Supplementary Fig. 5**). Lastly, half of these most active genera contained the capacity for nitrogen reduction via respiration or by dissimilatory nitrate reduction to ammonium (DNRA) (**Supplementary Fig. 6**). Together this encoding of both aerobic and anaerobic energy systems, and diverse electron donors across the many core, active taxa highlight the metabolic versatility harbored in surface waters.

Some critical river biogeochemical processes such as nitrification were represented by GROWdb MAGs but were not sampled in the top 25 most active genera. In surface waters, nitrification appeared to be catalyzed by bacteria, a finding consistent with taxonomy profiles from our unassembled reads where archaea were less than 3% of the relative abundance across samples (**Supplementary Fig. 7**). We identified one MAG within the *Nitrosomonas* genus with genes for ammonia oxidation (the first step in nitrification). We note this genome also included genes to produce the greenhouse gas nitrous oxide (N_2O), a finding consistent with other ammonia oxidizing bacteria[35].

Two other GROWdb MAGs contained genes for nitrite oxidation (the second step in nitrification) with taxonomy assignments to the *Nitrospira_D* genus and an unassigned species within the Palsa1315 genus of the Nitrospiraceae family. With these genomes being up to 95% complete, we infer comamomox is unlikely as these MAGs contained genes for nitrite oxidation but lacked genes for ammonia oxidation. These two GROW nitrite oxidizers were detected in 14-88% of the metatranscriptome samples, including detection of transcripts for the key protein in nitrite oxidation. Each of the three nitrifier MAGs contained genes for combating reactive oxygen species (superoxide dismutase, catalase, and/or peroxidase) and a photolyase gene involved in the repair of damage caused by exposure to ultraviolet light, all adaptations likely important in surface waters[36]. Overall, our findings uncover the nitrifier metabolic potential and expression in rivers, which are underrepresented in genomic databases compared to nitrifiers from soil and marine habitats.

While not core members, we also detected 17 Patescibacterial MAGs with at least on active gene, members of a phylum that currently lacks any pure culture representatives. There were 48 total Patescibacterial MAGs, all of which lacked the capacity for aerobic or anaerobic respiration and were inferred to be obligate fermenters, consistent with prior genomic reports from this phylum[37,38]. We verified the abundance patterns reported here were consistent with other river metagenome and amplicon-based studies[39,40], where these lineages accounted for up to 7% of the relative abundance in river surface water communities. It is possible that these obligately anaerobic members thrive in lower oxygen niches through endosymbiotic and biofilm lifestyles or via microsites of ephemeral oxygen depletion that occur

during dissolved organic matter decomposition[41,42]. In support of the latter, we observed that relative abundance and expression of Patescibacteria significantly decreased with river size (**Supplementary Fig. 8**), suggesting these obligate fermenters were more adapted to slower moving streams where the metabolic driver of aggregates is more sedimentarily associated[43].

The occurrence of antibiotics and antibiotic resistance genes (ARGs) in riverine systems has become a growing concern worldwide. As rivers flow with heavy antibiotic burdens, antibiotic resistance develops rapidly and disseminates into various environmental compartments[44]. Antibiotic production is also part of natural competition in these complex communities. We cataloged 1587 candidate ARGs recovered from 1135 (54.3%) MAGs in GROWdb, representing 25 different Phyla. Approximately 29% of the ARGs had evidence of expression in the metatranscriptome of one or more samples, and eleven samples had at least 20 ARGs with evidence of expression. Because our analysis was MAG-focused, these numbers may represent a floor on ARG prevalence in rivers, as they do not include plasmid-encoded ARGs. These candidate ARGs represent 25 broad antimicrobial resistance gene families as defined by the Comprehensive Antibiotic Resistance Database (CARD)[45]. Individual MAGs sometimes coded for ARGs from multiple AMR gene families and targeting multiple drug classes.

Interestingly, most (1219) candidate ARGs were homologs of proteins coded in glycopeptide resistance (*van*) gene clusters, which occurred in 955 distinct MAGs. Although glycopeptide antibiotics and self-protective resistance genes are often associated with gram-positive actinomycete producers[46,47], in the GROW data 65% of the MAGs containing *van* homologs were found in 23 other diverse non-Actinobacteriota phyla, including 140 MAGs with at least 2 unique *van* gene types. Given self-protective glycopeptide resistance gene clusters have been found adjacent to or as part of a variety of different antibiotic biosynthetic gene clusters in actinobacteria, this diversity suggests that rivers, like soils, might harbor both unique antibiotic biosynthetic gene clusters and antibiotic-resistance genes.

River microbiomes exhibit biogeographical patterns at the continental scale

One of the strengths of our experimental design was the spatial, chemical, and physical variables that accompanied our microbiome sampling, which allowed us to contextualize the factors driving microbial biogeography at the continental scale. Previous studies have done this at local or regional levels[7,48], and used taxonomy alone, but to our knowledge, this has yet to be done at the continental scale, nor incorporated functional gene-trait information. We hypothesized that river microbial communities exhibit biogeographical patterns at the continental scale of inquiry, and that these patterns would be predictable from hydrobiogeochemical, geographic, and land management factors. Every sample had a paired suite of more than 250 physical, chemical, and spatial variables (e.g., stream size, latitude, total nitrogen), which we could then use to see which were the strongest drivers of microbiome structure and expressed function (**Supplementary Table 1**).

Of all the river site variables examined (**Fig. 4A**), stream order, a numerical ranking of the relative size of streams that spans small headwater streams (order 1-3) to larger order rivers like the Mississippi river (order 8-12), was the most important controller of the microbiome composition. River size was more important than latitudinal position or total carbon which are often cited as controllers of microbiomes across other habitats[49,50]. Both metagenomes and metatranscriptomes were structured by stream order (Fig. 4A and 4B), providing evidence in favor of the river continuum concept, discussed in more detail in the next section. After stream order, expressed microbial functional profiles were also influenced by watershed temperature (both mean and maximum), area, and total runoff (**Fig. 4A**).

Given this relationship to temperature we sought to understand which functional traits and organisms were most contributing to these community level observations. Regression based modeling showed light driven metabolisms, followed by aerobic processes, were the most important variables predictive of mean and max watershed temperature (**Fig. 4C**). The most important organismal predictors of maximum watershed temperature were the core active lineages like *Methylopumilis*, UBA954, *Polynucleobacter*, and *Limnohabitans* that were actively transcribing genes for light harvesting metabolisms **(Fig. 3C)**. Our findings show that light harvesting metabolisms are critical to energy generation in rivers, and that temperature plays a defining role in the niches of these microorganisms.

These findings are consistent with reports from marine systems[51], hinting and an emerging rule set shared across aquatic microbiomes.

Beyond environmental factors we also observed that geographic position played a role in structuring river microbiomes. For instance, microbial community genomic membership was structured across ecoregions defined by Omernik level II classifications[52], with drier climate mixed grass influenced river microbiomes sharing similar microbial communities that were distinct from those derived from wet to sub-tropical regions (**Fig. 4D**). Similarly, hydrologic unit code (HUC), a classification system for watersheds in the United States shown in **Fig. 1C**, recognized distinct microbial communities from continental subregions (**Supplementary Fig. 9**). We note that the spatial structuring was not observed at the expressed functional level (**Supplementary Fig. 9**), indicating species changes are compensated by functional equivalence at this continental scale. This finding suggests taxonomic information may not be best suited for translation of microbiome content into management indicators, unless incorporated into an eco-regional framework as has been suggested for soil health indicators[53,54].

To use microbiota information as sentinels for monitoring human and environmental health in river systems, a greater understanding of bacterial community structure, function, and variability in lotic systems is required[23]. Particularly important is how different land management types may structure surface water microbiomes. Here we used a machine learning approach to understand river MAGs or gene content and expression associations with different land management regimes. We found that land use predicted metatranscriptomics richness (72% accuracy), with agriculture and urban influences being the most predictive variables (**Fig. 4E**). Together our findings show how river microbiomes are driven by multiple environmental factors, including local, watershed, regional, and continental factors. Analyses using GROWdb provide a new framework for the environmental factors and determinant mechanisms that shape riverine ecosystem communities.

Microbial gene expression provides evidence for the river continuum concept

The River Continuum Concept (RCC) provides a framework for integrating predictable and observable biological features of flowing water systems, and further characterizing how biodiversity changes along a river system[25]. Specifically, the RCC postulates that as rivers increase in size, the influences of terrestrial inputs will decrease. It also assumes, that biological richness will initially increase with stream order complexity due to maximum interface with the landscape, but then decrease along with river width and discharge. Support for the RCC extending from macro invertebrates to microbiota has been provided using decreased microbial 16S rRNA gene richness occurring across a stream order gradient in Toolik[55], Danube[56], Mississippi[23], and Amazon[19] rivers. As initial support for the RCC, our biogeographical analyses revealed that stream order was an important driver of microbial community genomic and transcriptomic patterns at the continental scale (**Fig. 4B**), providing an opportunity to interrogate diversity and functional trait dimensions along a river continuum at broader scales.

First, we were interested in how microbial richness at the metagenome and metatranscriptome level changed across the stream order gradient, and if these followed rules like 16S rRNA studies from single rivers. At the metagenome level, overall genome richness had a positive correlation with stream order, peaking at a stream order of six (**Fig. 4B**). At the metatranscriptome level, richness had a strong positive correlation with stream order and peaked at stream order 8, the highest stream order profiled by metatranscriptomics (**Fig. 4B**). Metagenome results were consistent with previous reports of the RCC where stream order peaks in mid-sized streams[23]. To our knowledge this is the first report of genome-resolved metatranscriptomics across rivers, suggesting that genome inferred transcriptional richness may be governed by different theoretical underpinnings than presence at the continental scale.

Our multi-omics data allowed us to examine the microbial functional gene changes across a river continuum. It is often cited in the RCC that smaller order streams are characterized by thick shore vegetation or overhanging trees that limit sunlight penetration and restrict phytoplankton primary production[25,57]. Consistent with this supposition, we observed statistically significant increase in light-driven microbial metabolisms when moving from smaller order streams to larger order rivers (**Fig. 4A**). Additionally, the RCC proposes the ratio of photosynthesis to respiration (P/R) increases in medium-

larger rivers but is decreased in the smallest and largest rivers due to light limitations from riparian vegetation occlusion and turbidity respectively. We used microbial gene expression to estimate P/R ratios, revealing the highest P/R ratio in rivers with stream orders of 6-8, providing tentative support for this concept. However, the robustness of this P/R indicator would need further evaluation in larger ordered rivers (e.g., 9-12) rivers, which are under sampled in this metatranscriptome dataset.

Based on the RCC, we also hypothesized gene expression would show that microbial carbon usage reflects decreasing impacts of terrestrial inputs with river size. To resolve changes in microbial metabolism across a stream order gradient, we defined carbon usage patterns based on microbial gene expression in GROWdb MAGs. Our findings show significant differences in expressed microbial carbon usage following stream order gradient (**Fig. 5A**). Specifically, we see gene transcripts for utilizing polymers, aromatics, and sugars are upregulated in small order streams, while methylotrophy gene transcripts, primarily from methanol oxidation, increased in larger order rivers. We consider methanol is likely autochthonous, derived from river phytoplankton biomass[58] or microbial metabolism of aromatic allochthonous plant litter[59,60]. Our findings show that the inferred microbial metabolisms related to carbon usage follow the expected decrease in impact of terrestrial inputs proposed by the RCC, but we acknowledge more research is needed to validate these new insights, especially from larger order rivers.

In summary, river systems have long been thought to be passive pipes, transporting water from terrestrial to marine systems. As a result, it was often regarded that microbiomes in these systems were a mixture of randomly assembled microorganisms with little predictive capability. Instead, we show river microbiomes and encoded functionalities are not haphazardly distributed but are instead structured by river size, ecological region, and land management regimes. We go on to provide the first gene functional support for the RCC and extend these findings across multiple river networks. We show that microbial richness both in terms of genome potential and expression, as well as expressed functional attributes follow RCC tenets and are molded by the physical-geomorphic environment. This application of GROWdb to the RCC bolsters previous microbial reports while also adding a new view of how microbial metabolism changes across river orders.

Conclusions

With increasing impacts of ongoing environmental change on microbial populations across ecosystems, a grand challenge is rapidly obtaining data that can generate transferable understanding linking microbes to ecosystem processes. Addressing this challenge requires samples across environmentally diverse settings, which can be prohibitive for a single research group to collect in a timely manner. Instead, to build GROWdb in a three-year period we used a crowdsourced approach to obtain samples from North American watersheds. We also ensured all samples were processed, sequenced, and analyzed using standardized methods to enable cross-site comparisons. Using this approach, we nearly doubled the number of samples from rivers in the sequence read archive (SRA), as well as tripled the amount of sequencing (Tb) for river microbiomes.

We sequenced 215 metagenomes and metatranscriptomes, sampling over 90% of the watersheds in the United States, with samples spanning different river sizes, geographic locations, and land management regimes. GROWdb is a comprehensive resource that includes a collection of over two thousand curated metagenome-assembled genomes, represented by 1,469 species spanning 27 phyla. As a database, GROWdb recruits an average of 52% of the reads across samples and up to 90% of reads in a single sample. This indicates that GROWdb increases mappability and reduces the computational requirements for expediting the translation of reads to functional content. Ultimately, GROWdb streamlines the exploitation of functional microbiome content, making microbiota derived insights more accessible across disciplinary boundaries.

To ensure data accessibility we provide four layers for user engagement with GROWdb. First, all reads and MAGs are publicly hosted on National Center for Biotechnology (NCBI), enabling transferability to resources (e.g., GTDB[61]) that pull and incorporate this content. Datasets underlying GROWdb are available and searchable through the free NMDC Data portal, linking other metadata types (e.g., metabolome[62]) to allow for interoperability, and broader context where available. Next, GROWdb

MAGs are available as an annotated genomic collection in the freely accessible KBase cyberinfrastructure. In these interfaces, users can access sample information, gene- and MAG-level annotations, profile functional summaries, and genome scale models in a point and click interface. Lastly, to aid in data exploration we have compressed the taxonomic and functional insights from GROWdb into a web interface, called GROWdb Explorer, allowing our users to rapidly profile microbial organismal and functional distributions across the data set. GROWdb version 1 is an expanding repository to incorporate and unify global river multi-omic data for the future.

Changing climate impacts the quality and quantity of water in US rivers. For example, precipitation intensity, surface runoff, flooding, fires, sea level rise, and droughts affect water resources with direct impacts to human health, agriculture, energy production, and ecosystem resiliency[63]. Additionally, two thirds of the drinking water in the United States is sourced from surface waters. Acting as first responders, microorganisms exacerbate or mitigate biogeochemical processes and can serve as predictors for water quality under current and future water scenarios. This study offers the most comprehensive genomic view of river microbiota to date, providing a platform for high resolution investigations of river hydrobiogeochemistry. The results presented here reveal core functionalities of river microbiomes, illuminating the key metabolisms and microorganisms conserved and transcriptionally active across river systems. We envision that this genomic infrastructure will pave the way for future developments in water quality monitoring, identifying biomarkers indicative of land use or water quality changes (e.g., pollution, acidification, and eutrophication). Collectively, GROWdb fills a major knowledge gap in the current understanding of microbial diversity and function in river ecosystems, observations that can integrated into predictive watershed scale models.

Methods

Sample collection through crowdsourcing and standardization of workflows

To build GROWdb, we used a network-of-networks approach to obtain samples from across United States rivers. A majority of our samples were derived from the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) consortium[64], which is designed to facilitate the development of transferable scientific understanding and mutual benefit across stakeholders[30,32]. The WHONDRS sampling itself was based on sending free sampling kits, along with standardized protocols, to interested researchers globally (sites outside of North America are not included in GROWdb). These researchers volunteered their time to collect samples and sent the samples back for processing using consistent methods to enable cross-site comparisons, interoperable data, and transferable understanding. Samples from the WHONDRS consortium contributed 44% of the metagenomes and all the metatranscriptomes in GROWdb.

Samples collected under the WHONDRS 2019 sampling campaign (**Supplementary Data 1**) were first described in Garayburu-Caruso, et al.[65] Briefly, at each sampling location, collaborators selected sampling sites within 100 m of a station that measured river discharge, height, or pressure. Within each site, approximately 1L of surface water was sampled using a 60 mL syringe and was filtered through a 0.22 μm sterivex filter (EMD Millipore). Filters were capped, filled with 3mL of RNAlater, and shipped to Pacific Northwest National Laboratory on blue ice within 24 h of collection. Surface water samples and filters were immediately frozen at −20 °C upon receiving for nucleic acid extraction, respectively.

A key component of this analysis was the standardization that occurred in data processing and analyses. DNA and RNA were co-extracted at single facility at Colorado State University. DNA and RNA were coextracted from filters at Colorado State University using ZymoBIOMICS DNA/RNA Miniprep Kit (Zymo Research Cat. # R2002) coupled with RNA Clean & Concentrator-5 (Zymo Research Cat. # R1013). Samples were eluted in 40 μL and stored at −20 °C until sequencing.

Similarly, a Community Sequencing Project (CSP) provided by the Joint Genome Institute ensured that sequencing protocols and methodologies were consistent across the project. All the metagenomes and 23% of the metatranascriptomes were provided by JGI, with the balance of metatranscriptomes processed at University of Colorado Anschutz using the same methods and kits as

specified by JGI. Lastly, sequence data processing for each sample was performed using identical methods, with the GROWdb standard operating procedures documented on GitHub[22]. Collectively, the use of crowdsourced approaches, JGI support, and standardized methodologies resulted in GROWdb, a compendium of river microbiome data, an endeavor that would not have been possible to execute by a single laboratory alone.

Acquisition of geospatial data

The watershed statistics for each sample were primarily obtained from the Environmental Protection Agency's StreamCat database[66] and the National Hydrography Plus Version 2 (NHDPlus V2) Dataset using the *nhdplusTool*s package[67] in R (R Core Team, 2023). StreamCat provides over 600 consistently computed watershed metrics for all waterbodies identified in the US Geological Survey (USGS)'s NHDPlusV2 geospatial framework, making it a suitable data source for the broad spectrum of sample locations in this study.

For watershed metrics that were not included in StreamCat (i.e., dominant Omernik Ecoregion, mean net primary production, and mean aridity index), we first delineated each sample's watershed using *nhdplusTools*, then utilized the *terra* package[68] to aggregate the additional datasets across each site's watershed accordingly. This approach is consistent with SteamCat's geospatial methodology.

Lastly, we collected streamflow data for sites that had a nearby stream gauge. For locations without an identified co-located stream gauge (WHONDRS typically co-located their sample sites with a stream gage), we identified USGS stream gauges within 10 kilometers up- or downstream of our sampling locations using the *dataRetrieval* and *nhdplusTools* packages. All stream gages were then manually verified for their applicability to each sampling site (e.g., verifying there were no dams between the site and the stream gage, a major confluence, etc.). See **Supplementary Table 1** for a complete list of datasets included in our analysis. The complete R workflow for this geospatial analysis can be found on GitHub (Willi & Ross, 2023).

Metagenomic assembly, binning, and annotation

Genomic DNA was prepared for metagenomic sequencing using Plate-based DNA library preparation on the PerkinElmer Sciclone NGS robotic liquid handling system at the Joint Genome Institute. Briefly, one nanogram of DNA was fragmented and adapter ligated using the Nextera XT kit (Illumina) and unique 8bp dual-index adapters (IDT, custom design). The ligated DNA fragments were enriched with 12 cycles of PCR and purified using Coastal Genomics Ranger high throughput agarose gel electrophoresis size selection to 450-600bp. The prepared libraries were sequenced using Illumina NovaSeq sequencer following a 2x150nt indexed run recipe.

Resulting fastq files were assembled and binned using the GROWdb pipelines[22]. Briefly, three assemblies were performed on each set of fastq files and binned separately: (1) Read trimming with sickle (v1.33)[69], assembly with megahit (v1.2.9)[70], and binning with metabat2[71] (2.12.1) (2) Read trimming with sickle (v1.33)[69], random filtering to 25% of reads, assembly with idba-ud[72] (1.1.0), and binning with metabat2[71] (2.12.1) (3) Bins derived from the JGI-IMG pipeline[73] were downloaded. All resulting bins were assessed for quality using checkM[74] (v1.1.2) and medium and high-quality MAGs with >50% completion and <10% contamination were retained. The resulting 3,284 MAGs across all samples and assemblies were dereplicated at 99% identity using dRep[75] (v2.6.2) to obtain the dereplicated first version of the GROW database (n=2,093 MAGs). MAG taxonomy was assigned using GTDB-tk[61] (v2.1.1, r207) and annotated using DRAM[76].

To quantify MAG relative abundance across samples, trimmed metagenomic reads were mapped to the dereplicated MAG set using Bowtie2[77] and output as sam files which were then converted to sorted bam files using samtools. Sorted bam files were then filtered to paired reads only with a 95% identity match using reformat.sh. To obtain mean coverage for each MAG, we used CoverM[78] (-m trimmed_mean). The mean coverage table was then filtered to MAGs that had at least 60% coverage across a MAG with at least 3X coverage within a sample, using additional CoverM[78] outputs (-m

relative_abundance –min-covered-fraction 0.6 and -m reads_per_base, respectively). CoverM outputs were merged in R, with script available on the GROWdb github[22].

Metatranscriptomic mapping and analysis

RNA was prepared from metatranscriptome sequencing according to JGI established protocols. Briefly, rRNA was removed from 10 ng of total RNA using Qiagen FastSelect probe sets for bacterial, yeast, and plant rRNA depletion (Qiagen) with RNA blocking oligo technology. The fragmented and rRNA-depleted RNA was reverse transcribed to create first strand cDNA using Illumina TruSeq Stranded mRNA Library prep kit (Illumina) followed by second strand cDNA synthesis which incorporates dUTP to quench the second strand during amplification. The double stranded cDNA fragments were then A-tailed and ligated to JGI dual indexed Y-adapters, followed with an enrichment of the library by 13 cycles of PCR. The prepared libraries were quantified using KAPA Biosystems' next-generation sequencing library qPCR kit and run on a Roche LightCycler 480 real-time PCR instrument. Sequencing of the flowcell was performed on the Illumina NovaSeq sequencer following a 2x150nt indexed run recipe.

Resulting fastq files were mapped via Bowtie2[77] (-D 10 -R 2 -N 1 -L 22 -i S,0,2.50) to the dereplicated GROWdb. Sam files were transformed to bam files using samtools, filtered to 97% id using reformat.sh and name sorted using samtools. Transcripts were counted each gene with feature-counts[79]. Counts were transformed to geTMM (gene length corrected trimmed mean of M-values) in R using edgeR package[80]. Genes and bins were considered if they were expressed in 10% of samples, with bins level calculations having an additional requirement to express at least 20 genes.

Microbial metabolism trait and carbon usage classification

To classify carbon use across GROWdb and the metatranscriptomes, we curated the metabolism assignments made by DRAM using rulesets to assign genomes to functional guilds (Supplementary Fig. 3). From the DRAM output, we further assigned genomes as capable of carbon fixation if they encoded >70% of one of six seven carbon fixation pathways. We then assigned each MAG in each river metatranscriptome as a photoautotroph, photoheterotroph, chemolithoautotrophy, heterotroph, or mixotroph by assessing the gene expression in that system. We then focused in on carbon use in genomes identified for heterotrophy and assigned carbon gene expression into the following categories: polymer, sugar, aromatic compound, methanotrophy, methylotrophy, short chain fatty acid utilization, and carbon monoxide utilization. Carbon usage curation scripts are available on the GROWdb github[22].

Phylogenetic analyses were performed on genes annotated as respiratory nitrate reductase, nitrite oxidoreductase (nxr), ammonia monooxygenase (amo), or methane monooxygenase (pmo) to better assign the nitrogen cycling capabilities of GROW MAGs. Specifically, nxr/nar and pmo/amo reference sequences were downloaded[81–83]. Each set of reference sequences were combined with amino acid sequences of homologs from the GROWdb, aligned separately using MUSCLE (v3.8.31), and run through an in-house script for generating phylogenetic trees[84]. This resulted in two phylogenies, one for nxr/nar and one for pmo/amo, which were used to classify gene homologs in the MAG database.

P/R ratios were defined by the ratio of expression of light driven energy metabolisms (aerobic photosynthesis, anaerobic photosynthesis, and photorhodopsins) divided by respiration metabolisms (aerobic respiration and microaerophilic respiration).

Antibiotic Resistance Gene Classification

For *in silico* predictions of antimicrobial resistance genes (ARGs) GROWdb predicted proteins were searched for homology to proteins in the Comprehensive Antibiotic Resistance Database (CARD; v 3.2.7, downloaded June 2023) using the Resistance Gene Identifier (RGI; v 6.0.2)[45]. RGI was run locally in protein input mode with distributed input and default parameters, and with the “include loose” option. However, the final list of candidate ARGs analyzed here only includes proteins identified by RGI as “Perfect” or “Strict” hits, and only includes protein homolog models (i.e., no protein variant models were included in the analysis).

Sequence Read Archive analysis

To analyze the distribution of species recovered by GROW across public datasets, the Sandpiper database (https://sandpiper.qut.edu.au) was used as a basis. At the time of analysis, it contained metagenomes that were publicly available on Dec 15, 2021 (Woodcroft et. al., unpublished). Reanalysis of these datasets was carried out with SingleM 1.0.0beta7 (Woodcroft et. al., unpublished, https://github.com/wwood/singlem). The "supplement" subcommand was first used to add 95% ANI dereplicated GROW MAGs to the SingleM reference metapackage built with GTDB RS07-207 (DOI: 10.5281/zenodo.7582579). The "renew" subcommand was then used to reanalyze all metagenomes present in the Sandpiper database, outputting a taxonomic profile, detailing the species and unclassified lineages in each metagenome, together with their relative abundance.

To search for public metagenomes where GROW MAGs were present, taxonomic profiles of metagenomes containing species that had an associated GROW MAG (either novel or already represented in GTDB) were further analyzed. To reduce the incidence of false identification, we required at least 2 species represented by a GROW MAG to be present and the combined relative abundance of these species to be >1%. Metadata of metagenomes containing GROW MAGs were gathered using Kingfisher "annotate" (https://github.com/wwood/kingfisher-download).

Surface water geochemical characterization

Geochemical data collected under the WHONDRS 2019 sampling campaign are available ESS-DIVE, with methods described[62].

References

1. Villa, J. A. *et al.* Methane and nitrous oxide porewater concentrations and surface fluxes of a regulated river. *Sci. Total Environ.* **715**, 136920 (2020).
2. Raymond, P. A., Oh, N.-H., Turner, R. E. & Broussard, W. Anthropogenically enhanced fluxes of water and carbon from the Mississippi River. *Nature* **451**, 449–452 (2008).
3. Qu, B. *et al.* Greenhouse gases emissions in rivers of the Tibetan Plateau. *Sci. Rep.* **7**, 16573 (2017).
4. Aufdenkampe, A. K. *et al.* Riverine coupling of biogeochemical cycles between land, oceans, and atmosphere. *Front. Ecol. Environ.* **9**, 53–60 (2011).
5. Anderson, E. P. *et al.* Understanding rivers and their social relations: A critical step to advance environmental water management. *WIREs Water* **6**, e1381 (2019).
6. Alexander, R. B., Boyer, E. W., Smith, R. A., Schwarz, G. E. & Moore, R. B. The role of headwater streams in downstream water quality. *Journal of the American Water Resources Association* vol. 43 19 (2007).

7. Mishra, A., Alnahit, A. & Campbell, B. Impact of land uses, drought, flood, wildfire, and cascading events on water quality and microbial communities: A review and analysis. *J. Hydrol.* **596**, 125707 (2021).

8. Butman, D. & Raymond, P. A. Significant efflux of carbon dioxide from streams and rivers in the United States. *Nat. Geosci.* **4**, 839–842 (2011).

9. Kroeze, C., Dumont, E. & Seitzinger, S. P. New estimates of global emissions of N2O from rivers and estuaries. *Environ. Sci.* **2**, 159–165 (2005).

10. Rodríguez-Ramos, J. A. *et al.* Genome-Resolved Metaproteomics Decodes the Microbial and Viral Contributions to Coupled Carbon and Nitrogen Cycling in River Sediments. *mSystems* **7**, e00516-22 (2022).

11. Boddicker, A. M. & Mosier, A. C. Genomic profiling of four cultivated Candidatus Nitrotoga spp. predicts broad metabolic potential and environmental distribution. *ISME J.* **12**, 2864–2882 (2018).

12. Liu, S. *et al.* Comammox Nitrospira within the Yangtze River continuum: community, biogeography, and ecological drivers. *ISME J.* **14**, 2488–2504 (2020).

13. Chen, L., Li, F., Huang, W., Li, Z. & Chen, M. Evaluation of Mercury Transformation and Benthic Organisms Uptake in a Creek Sediment of Pearl River Estuary, China. *Water* **11**, 1308 (2019).

14. Ghosh, D., Ghosh, A. & Bhadury, P. Arsenic through aquatic trophic levels: effects, transformations and biomagnification—a concise review. *Geosci. Lett.* **9**, 20 (2022).

15. Stegen, J. C. *et al.* Groundwater–surface water mixing shifts ecological assembly processes and stimulates organic carbon turnover. *Nat. Commun.* **7**, 11237 (2016).

16. Read, D. S. *et al.* Catchment-scale biogeography of riverine bacterioplankton. *ISME J.* **9**, 516–526 (2015).

17. Savio, D. *et al.* Bacterial diversity along a 2600 km river continuum. *Environ. Microbiol.* **17**, 4994–5007 (2015).

18. Ghai, R. *et al.* Metagenomics of the Water Column in the Pristine Upper Course of the Amazon River. *PLOS ONE* **6**, e23785 (2011).

19. Satinsky, B. M. *et al.* Metagenomic and metatranscriptomic inventories of the lower Amazon River, May 2011. *Microbiome* **3**, 39 (2015).

20. Arora, B. *et al.* Building Cross-Site and Cross-Network collaborations in critical zone science. *J. Hydrol.* **618**, 129248 (2023).

21. Borton, M. *et al.* GROWdb US River Systems - Samples. (2022) doi:10.25982/109073.30/1895615.

22. Borton, M. A. Genome Resolved Open Watersheds database (GROWdb). (2023).

23. Henson, M. W. *et al.* Nutrient dynamics and stream order influence microbial community patterns along a 2914 kilometer transect of the Mississippi River. *Limnol. Oceanogr.* **63**, 1837–1855 (2018).

24. Doretto, A., Piano, E. & Larson, C. E. The River Continuum Concept: lessons from the past and perspectives for the future. *Can. J. Fish. Aquat. Sci.* **77**, 1853–1864 (2020).

25. Vannote, R. L., Minshall, G. W., Cummins, K. W., Sedell, J. R. & Cushing, C. E. The River Continuum Concept. *Can. J. Fish. Aquat. Sci.* **37**, 130–137 (1980).

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

27. URycki, D. R., Good, S. P., Crump, B. C., Chadwick, J. & Jones, G. D. River Microbiome Composition Reflects Macroscale Climatic and Geomorphic Differences in Headwater Streams. *Front. Water* **2**, (2020).

28. Payne, J. T., Millar, J. J., Jackson, C. R. & Ochs, C. A. Patterns of variation in diversity of the Mississippi river microbiome over 1,300 kilometers. *PLOS ONE* **12**, e0174890 (2017).

29. Katz, K. *et al.* The Sequence Read Archive: a decade more of explosive growth. *Nucleic Acids Res.* **50**, D387–D390 (2022).

30. Wilkinson, M. D. *et al.* The FAIR Guiding Principles for scientific data management and stewardship. *Sci. Data* **3**, 160018 (2016).

31. Eloe-Fadrosh, E. A. *et al.* The National Microbiome Data Collaborative Data Portal: an integrated multi-omics microbiome data resource. *Nucleic Acids Res.* **50**, D828–D836 (2022).

32. Goldman, A. E., Emani, S. R., Pérez-Angel, L. C., Rodríguez-Ramos, J. A. & Stegen, J. C. Integrated, Coordinated, Open, and Networked (ICON) Science to Advance the Geosciences: Introduction and Synthesis of a Special Collection of Commentary Articles. *Earth Space Sci.* **9**, e2021EA002099 (2022).

33. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A Guide to the Natural History of Freshwater Lake Bacteria. *Microbiol. Mol. Biol. Rev.* **75**, 14–49 (2011).

34. Jezbera, J., Sharma, A. K., Brandt, U., Doolittle, W. F. & Hahn, M. W. ‘Candidatus Planktophila limnetica’, an actinobacterium representing one of the most numerically important taxa in freshwater bacterioplankton. *Int. J. Syst. Evol. Microbiol.* **59**, 2864–2869 (2009).

35. Stein, L. Y. Insights into the physiology of ammonia-oxidizing microorganisms. *Curr. Opin. Chem. Biol.* **49**, 9–15 (2019).

36. Haney, A. M., Sanfilippo, J. E., Garczarek, L., Partensky, F. & Kehoe, D. M. Multiple Photolyases Protect the Marine Cyanobacterium Synechococcus from Ultraviolet Radiation. *mBio* **13**, e01511-22 (2022).

37. Tian, R. *et al.* Small and mighty: adaptation of superphylum Patescibacteria to groundwater environment drives their genome simplicity. *Microbiome* **8**, 51 (2020).

38. Wrighton, K. C. *et al.* Fermentation, Hydrogen, and Sulfur Metabolism in Multiple Uncultivated Bacterial Phyla. *Science* **337**, 1661–1665 (2012).

39. Lian, Y., Zhen, L., Chen, X., Li, Y. & Li, X. Microbial biomarkers as indication of dynamic and heterogeneous urban water environments. *Environ. Sci. Pollut. Res.* (2022) doi:10.1007/s11356-022-24539-8.

40. Regina, A. L. A. *et al.* A watershed impacted by anthropogenic activities: Microbial community alterations and reservoir of antimicrobial resistance genes. *Sci. Total Environ.* **793**, 148552 (2021).

41. Ploug, H., Kühl, M. & BUCHHOLZCLEVEN, B. Anoxic aggregates - An ephemeral phenomenon in the pelagic environment? *Aquat. Microb. Ecol.* **13**, 285–294 (1997).

42. Böckelmann, U., Manz, W., Neu, T. R. & Szewzyk, U. Characterization of the microbial community of lotic organic aggregates (‘river snow’) in the Elbe River of Germany by cultivation and molecular methods. *FEMS Microbiol. Ecol.* **33**, 157–170 (2000).

43. Battin, T. J. *et al.* Biophysical controls on organic carbon fluxes in fluvial networks. *Nat. Geosci.* **1**, 95–100 (2008).

44. Jia, S. *et al.* Fate of antibiotic resistance genes and their associations with bacterial community in livestock breeding wastewater and its receiving river water. *Water Res.* **124**, 259–268 (2017).

45. Alcock, B. P. *et al.* CARD 2023: expanded curation, support for machine learning, and resistome prediction at the Comprehensive Antibiotic Resistance Database. *Nucleic Acids Res.* **51**, D690–D699 (2023).

46. Waglechner, N., McArthur, A. G. & Wright, G. D. Phylogenetic reconciliation reveals the natural history of glycopeptide antibiotic biosynthesis and resistance. *Nat. Microbiol.* **4**, 1862–1871 (2019).

47. Yushchuk, O., Binda, E. & Marinelli, F. Glycopeptide Antibiotic Resistance Genes: Distribution and Function in the Producer Actinomycetes. *Front. Microbiol.* **11**, (2020).

48. Pan, Y., Xie, J., Yan, W., Zhang, T. C. & Chen, C. Response of microbial community to different land-use types, nutrients and heavy metals in urban river sediment. *J. Environ. Manage.* **321**, 115855 (2022).

49. Dickey, J. R. *et al.* The Utility of Macroecological Rules for Microbial Biogeography. *Front. Ecol. Evol.* **9**, (2021).

50. Smith, L. C. *et al.* Large-scale drivers of relationships between soil microbial properties and organic carbon across Europe. *Glob. Ecol. Biogeogr.* **30**, 2070–2083 (2021).

51. DeLong, E. F. Microbial community genomics in the ocean. *Nat. Rev. Microbiol.* **3**, 459–469 (2005).

52. Omernik, J. M. Ecoregions of the Conterminous United States. *Ann. Assoc. Am. Geogr.* **77**, 118–125 (1987).

53. Smith, J. R., Hendershot, J. N., Nova, N. & Daily, G. C. The biogeography of ecoregions: Descriptive power across regions and taxa. *J. Biogeogr.* **47**, 1413–1426 (2020).

54. Fine, A. K., van Es, H. M. & Schindelbeck, R. R. Statistics, Scoring Functions, and Regional Analysis of a Comprehensive Soil Health Database. *Soil Sci. Soc. Am. J.* **81**, 589–601 (2017).

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

56. Mathias, C. B., Kirschner, A. & Velimirov, B. Seasonal variations of virus abundance and viral control of the bacterial production in a backwater system of the danube river. *Appl. Environ. Microbiol.* **61**, 3734–3740 (1995).

57. Maiolini, B. & Bruno, M. C. The River Continuum Concept revisited: Lessons from the Alps. (2023).

58. Mincer, T. J. & Aicher, A. C. Methanol Production by a Broad Phylogenetic Array of Marine Phytoplankton. *PLoS ONE* **11**, e0150820 (2016).

59. McInerney, M. J. *et al.* Physiology, Ecology, Phylogeny, and Genomics of Microorganisms Capable of Syntrophic Metabolism. *Ann. N. Y. Acad. Sci.* **1125**, 58–72 (2008).

60. Schink, B. & Zeikus, J. G. Microbial methanol formation: A major end product of pectin metabolism. *Curr. Microbiol.* **4**, 387–389 (1980).

61. Chaumeil, P.-A., Mussig, A. J., Hugenholtz, P. & Parks, D. H. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**, 1925–1927 (2020).

62. Toyoda, J, Goldman, A.E., Chu, R.K., Danczak, R.E., Daly, R.A., Garayburu-Caruso, V.A., Graham, E.B., Lin, X. Moran, J.J., Ren, H., Renteria, L. Resch, C.T., Tfaily, M., Tolic N., Torgeson, J.M., Wells J., Wrighton, K.C., Stegen, J.C. WHONDRS Consortium T (2020):

WHONDRS Summer 2019 Sampling Campaign: Global River Corridor Surface Water FTICR-MS, NPOC, and Stable Isotopes. (2020) doi:10.15485/1603775.

63. Gudmundsson, L. *et al.* Globally observed trends in mean and extreme river flow attributed to climate change. *Science* **371**, 1159–1162 (2021).

64. Stegen, J. C. & Goldman, A. E. WHONDRS: a Community Resource for Studying Dynamic River Corridors. *mSystems* **3**, (2018).

65. Garayburu-Caruso, V. A. *et al.* Using Community Science to Reveal the Global Chemogeography of River Metabolomes. *Metabolites* **10**, 518 (2020).

66. Hill, R. A., Weber, M. H., Leibowitz, S. G., Olsen, A. R. & Thornbrugh, D. J. The Stream-Catchment (StreamCat) Dataset: A Database of Watershed Metrics for the Conterminous United States. *JAWRA J. Am. Water Resour. Assoc.* **52**, 120–128 (2016).

67. Blodgett, D., Johnson, J. M. & Bock, A. Generating a reference flow network with improved connectivity to support durable data integration and reproducibility in the coterminous US. *Environ. Model. Softw.* **165**, 105726 (2023).

68. Hijmans, Robert J., et al. Package 'terra'. *Maint. Vienna Austria* (2022).

69. najoshi. sickle - A windowed adaptive trimming tool for FASTQ files using quality. (2023).

70. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).

71. Kang, D. D. *et al.* MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**, e7359 (2019).

72. IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth | Bioinformatics | Oxford Academic. https://academic.oup.com/bioinformatics/article/28/11/1420/266973.

73. Clum, A. *et al.* DOE JGI Metagenome Workflow. *mSystems* **6**, 10.1128/msystems.00804-20 (2021).

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

75. Olm, M. R., Brown, C. T., Brooks, B. & Banfield, J. F. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J.* **11**, 2864–2868 (2017).

76. Shaffer, M. *et al.* DRAM for distilling microbial metabolism to automate the curation of microbiome function. *Nucleic Acids Res.* **48**, 8883–8900 (2020).

77. Langmead, B. & Salzberg, S. L. Fast gapped-read alignment with Bowtie 2. *Nat. Methods* **9**, 357–359 (2012).

78. Woodcroft, B. J. CoverM. (2023).

79. Liao, Y., Smyth, G. K. & Shi, W. featureCounts: an efficient general purpose program for assigning sequence reads to genomic features. *Bioinformatics* **30**, 923–930 (2014).

80. Robinson, M. D., McCarthy, D. J. & Smyth, G. K. edgeR: a Bioconductor package for differential expression analysis of digital gene expression data. *Bioinformatics* **26**, 139–140 (2010).

81. Tavormina, P. L., Orphan, V. J., Kalyuzhnaya, M. G., Jetten, M. S. M. & Klotz, M. G. A novel family of functional operons encoding methane/ammonia monooxygenase-related proteins in gammaproteobacterial methanotrophs. *Environ. Microbiol. Rep.* **3**, 91–100 (2011).

82. Rochman, F. F. *et al.* Novel copper-containing membrane monooxygenases (CuMMOs) encoded by alkane-utilizing Betaproteobacteria. *ISME J.* **14**, 714–726 (2020).

83. Daims, H. *et al.* Complete nitrification by Nitrospira bacteria. *Nature* **528**, 504–509 (2015).

84. WrightonLabCSU. *GitHub* https://github.com/WrightonLabCSU.

85. Wood-Charlson, E. M. *et al.* The National Microbiome Data Collaborative: enabling microbiome science. *Nat. Rev. Microbiol.* **18**, 313–314 (2020).

86. Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **36**, 566–569 (2018).

Data accessibility

The data underlying GROWdb are accessible across various platforms to ensure all levels of data structure are widely available. First, all reads and MAGs are publicly hosted on National Center for Biotechnology (NCBI) under Bioproject PRJNA946291. Second, all data presented here including MAG annotations, phylogenetic tree files, antibiotic resistance gene database files, and expression data tables are available in Zenodo (https://doi.org/10.5281/zenodo.8173287).

Beyond the flat database files listed above, our aim for GROWdb was to maximize data use by making the data available in searchable and interactive platforms including the National Microbiome Data Collaborative (NMDC)[31,85] data portal, the Department of Energy's Systems Biology Knowledgebase (KBase)[86], and a GROW specific user interface released here, GROWdb Explorer. Each platform provides different ways to interact with GROWdb:

- *NMDC* GROWdb was a flagship project for the newly formed NMDC. Specifically, individual GROWdb datasets (metagenomes, metatranscriptomes, etc) are easily accessible and searchable through the NMDC data portal (https://data.microbiomedata.org/), where they are systematically connected to each other and to a rich suite of sample information and standard analysis results, following Findable, Accessible, Interoperable, and Reusable (FAIR) data practices[30].
- *KBase* GROWdb is publicly available collection within KBase[86], with samples, MAGs, and corresponding genome scale metabolic models found in the KBase narrative structure (https://doi.org/10.25982/109073.30/1895615). Access within KBase allows for immediate access and reuse of data, including comparison to private data using KBase's 500+ analysis tools.

- *GROWdb Explorer* is a graphical user interface built through the Colorado State University Geospatial Centroid (https://geocentroid.shinyapps.io/GROWdatabase/), allowing users to search and graph microbial and spatial data simultaneously. Here we the microbial data was distilled into functional gene information, so that biogeochemical contributions and the microorganisms catalyzing them can be assessed and visualized rapidly across the dataset.

In summary, GROWdb represents the first publicly available genome collection from rivers and offers data that can be leveraged across microbiome studies. GROWdb is an expanding repository to incorporate and unify global river multi-omic data for the future.

Funding

This work was partially supported by awards from U.S. Department of Energy (DOE) Office of Science, Office of Biological and Environmental Research (BER) grants DE-SC0023084 (MAB, BBM, KCW) and DE-SC0021350 (MAB, DMS, CSH, CSM, KCW). A portion of this work was also performed by MAB under a subcontract to KCW from the River Corridor Science Focus Area (RCSFA) at Pacific Northwest National Laboratory (PNNL) and funded by the DOE BER Environmental System Science (ESS) Program. PNNL is operated by Battelle Memorial Institute for the DOE under Contract No. DE-AC05-76RL01830. WHONDRS efforts described in this manuscript, JCS, and AEG were also funded under the RCSFA at PNNL by DOE BER ESS. Metagenomic and metatranscriptomic sequencing was performed at the Joint Genome Institute under a Community Science Program and the University of Colorado Anschutz's Genomics Shared Resource. The work (proposal: 10.46936/10.25585/60001289) conducted by the U.S. Department of Energy Joint Genome Institute (https://ror.org/04xm1d337), a DOE Office of Science User Facility, is supported by the Office of Science of the U.S. Department of Energy operated under Contract No. DE-AC02-05CH11231.Work conducted at the Genomics Shared Resource was supported by the Cancer Center Support Grant (P30CA046934). The work conducted by the National Microbiome Data Collaborative (https://ror.org/05cwx3318) is supported by the Genomic Science Program in the U.S. Department of Energy, Office of Science, Office of Biological and Environmental Research (BER) under contract numbers DE-AC02-05CH11231 (LBNL), 89233218CNA000001 (LANL), and DE-AC05-76RL01830 (PNNL).

Acknowledgements

Samples were sequenced and processed as a part of the Genome Resolved Open Watersheds database (GROWdb) effort to sequence global watersheds. A portion of samples and data used in this manuscript were generated as a part of the WHONDRS global crowdsourced Summer 2019 Sampling (S19S) and we thank those that participated in the design and implementation of that effort. We also thank Tyson Claffey and Richard Wolfe for Colorado State University server management and Zach Crockett at KBase for generation of sample set Digital Object Identifier.

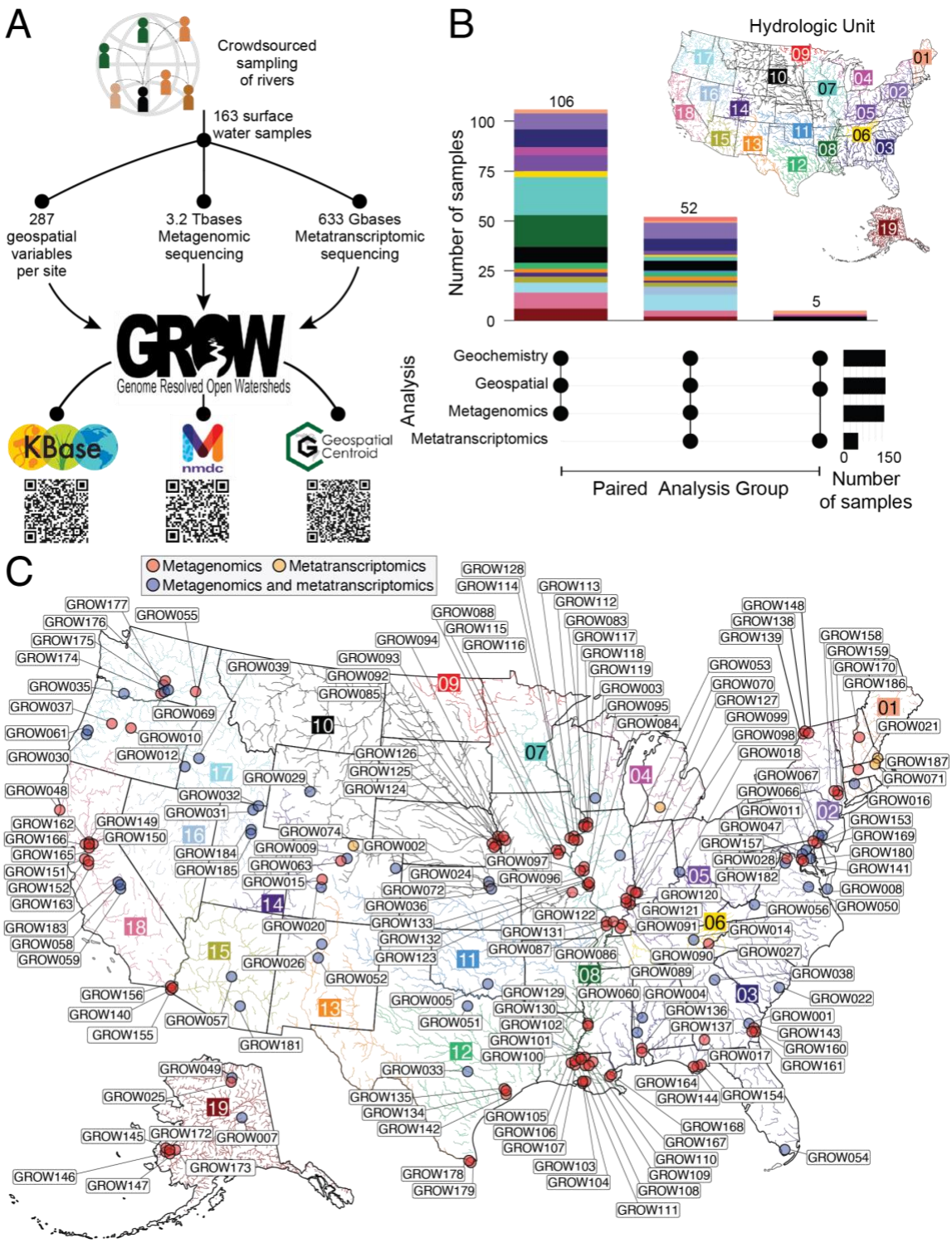


Figure 1. Distributed sampling and sequencing of rivers across the United States enabled the construction of the first version of Genome Resolved Open Watersheds (GROW). A) Workflow denoting the number of samples and resulting datasets made up of geospatial and microbiome (metagenomics, metatranscriptomics) data. GROWdb data are accessible through KBase, NMDC, and the Geospatial Centroid, with QR codes directing to online data location. B) Plot showing the number of samples with paired data types (denoted as filled black circles below) as stacked bars colored by hydrologic unit, and number of samples per analysis as black solid bars. C) GROW sampling across the

United States with points marking a sampling location, color-coded by the microbiome analysis performed (MetaG, red; MetaT, yellow; paired MetaG and MetaT, blue). Boxed numbers and corresponding river color indicate hydrologic unit (HUC-2). Hawaii not shown.

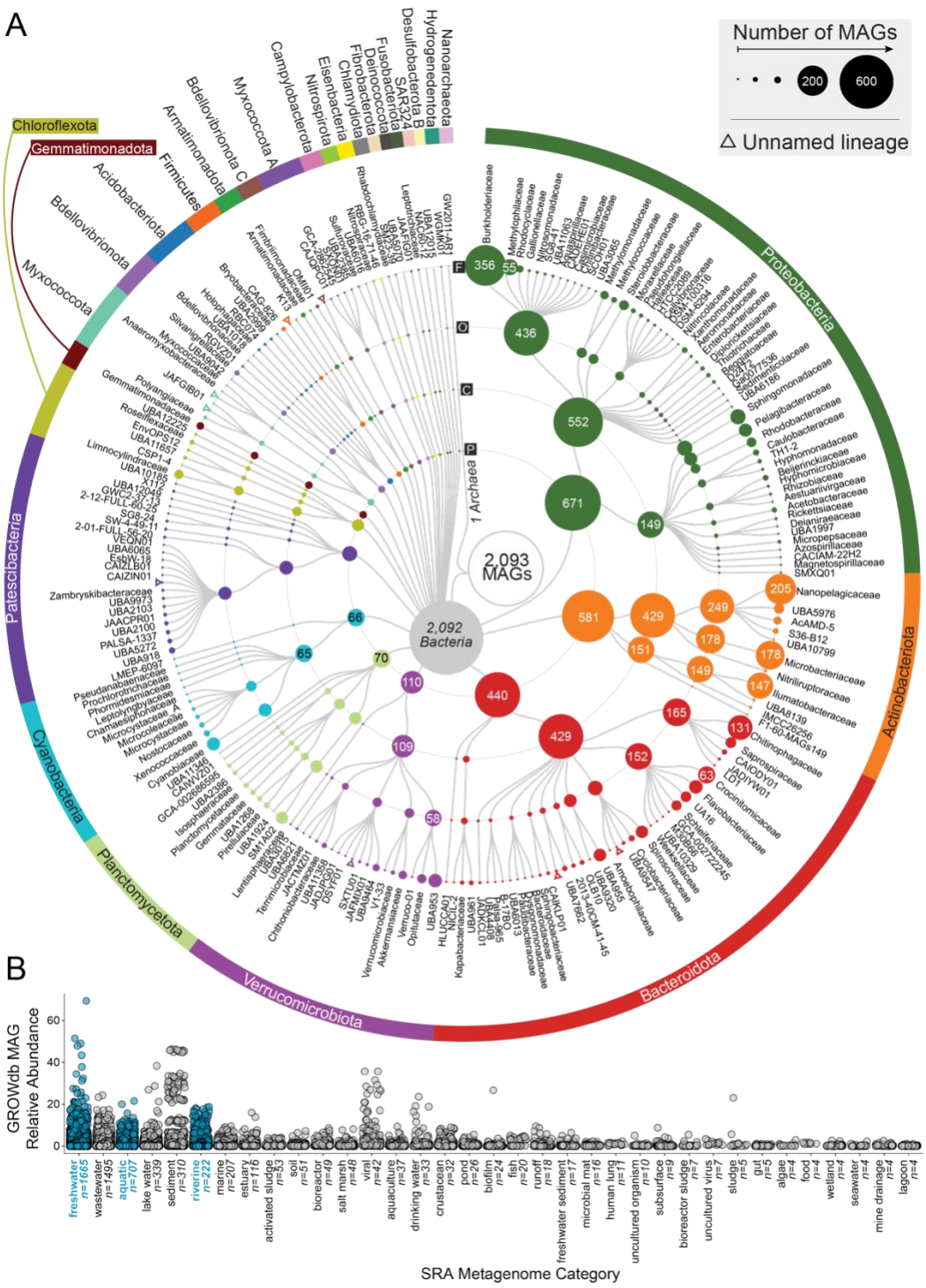


Figure 2: Taxonomic diversity of 2,093 unique surface water metagenome assembled genomes (MAGs) in GROWdb. A) Cladogram shows taxonomy of GROWdb MAGs with each sequential ring noting taxonomy level (Phylum, P; Class, C; Order, O; Family, F) as defined by GTDB-Tk. Circle size indicates the number of genomes within a given taxonomy level and is further noted by MAG number inside the circle with greater than 50 MAGs shown. Colors highlight Phylum level taxonomy denoted by the outermost ring. Open triangles represent unnamed lineages within a particular level of taxonomy. B) One dimensional scatterplot displays the environments GROWdb MAGs were detected in across the

Sequence Read Archive (SRA, n=266,764 metagenomes) with each point representing a single MAG. Environments are ordered by number of metagenomes GROWdb MAGs were detected in from left to right, with the number of metagenomes also noted along the x-axis. Freshwater related environments are highlighted in blue.

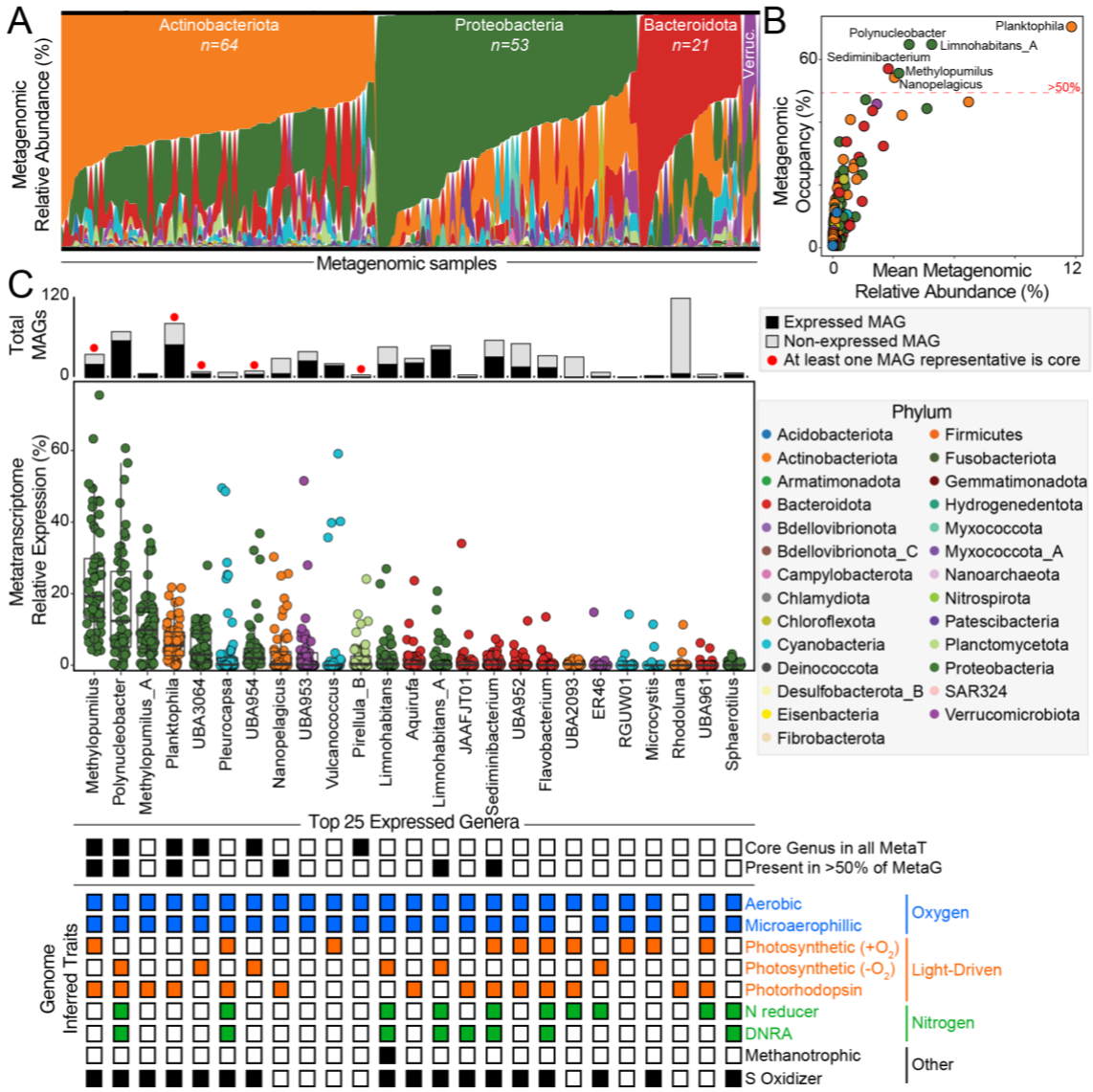


Figure 3: Core lineages and functions across microbiomes. A) Phyla dynamics across metagenomes are shown, with Actinobacteriota, Proteobacteria, Bacteroidota, and Verrucomicrobiota (Verruc.) being the most dominant phyla across samples. B) Scatter plot highlights metagenomic relative abundance versus metagenomic occupancy (percent of metagenomes a genera was present in), with points representing each genera in GROWdb and colored by Phylum. Genera in detected in more than 50% of samples (red dashed line) are named. C) The top 25 most transcribed (highest metatranscriptomic expression) genera are shown by boxplots, with each point representing a single metatranscriptome. Stacked bar chart above boxplots indicates the number of MAGs in GROWdb within each genus and colored by detection in metatranscriptomes (black, expressed; grey, non-expressed). A red circle above bar indicates that one of the genomes was core across metatranscriptomes, or expressed in every sample. For each of the top 25 expressed genera, detection across metatranscriptomes (core, 100% of

metatranscriptomes) and metagenomes (>50% of metagenomes) is displayed by black boxes. Below, inferred genomic potential of each genera is indicated, including aerobic respiration (blue), light-driven energy metabolism (orange), nitrogen metabolism (green), and other metabolisms (e.g., methanotrophy and sulfur oxidation, black).

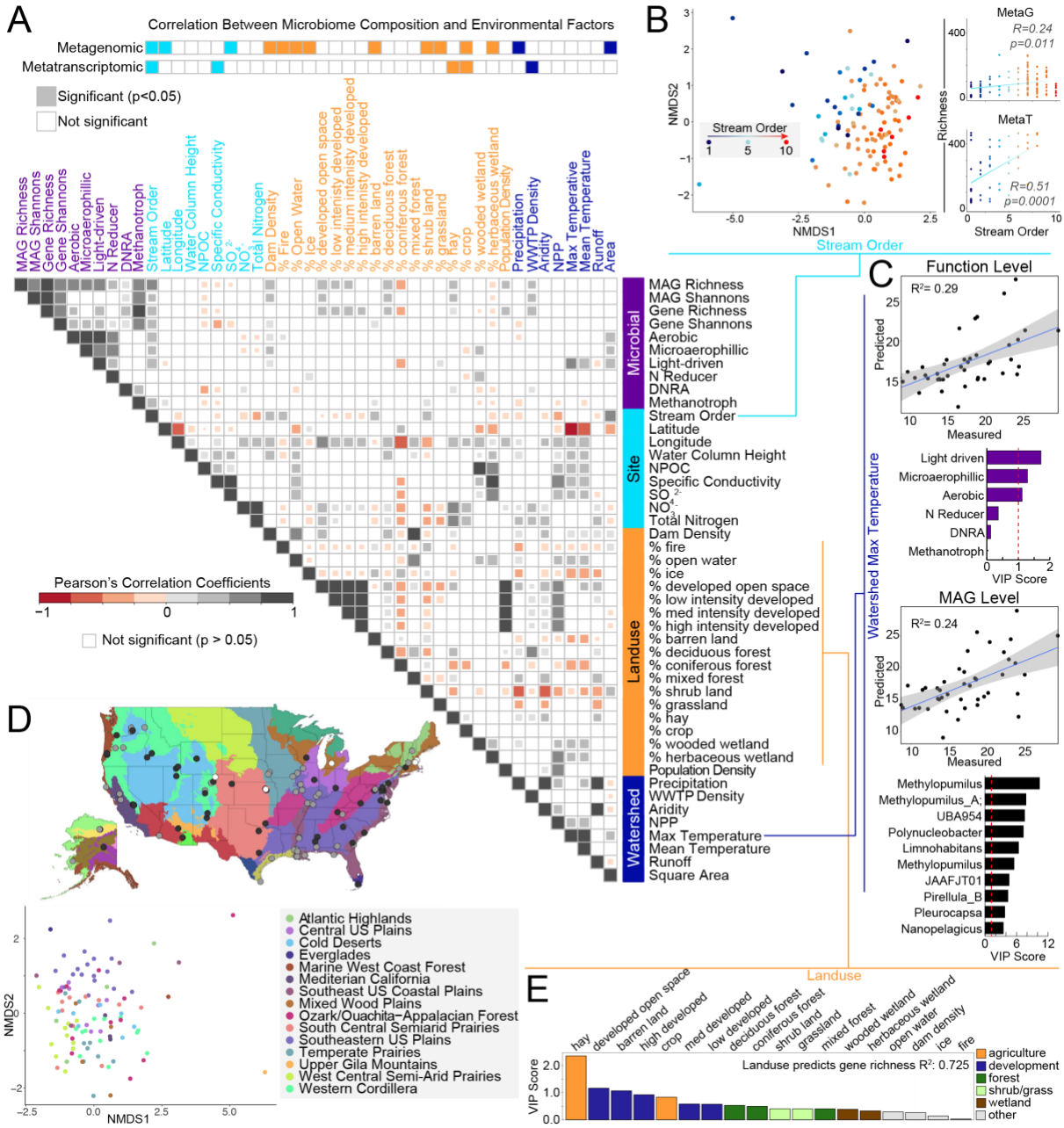


Figure 4: Patterns and drivers of river microbiome composition and function. A) Metagenomic and metatranscriptomic composition, function, and diversity were related to 36 selected geospatial variables using Mantel tests (top, two rows) along with pairwise comparisons using Pearson's correlation. Variables are colored by category including microbial (purple), site or local (light blue), land use (orange) and watershed metrics (dark blue), which are carried through the rest of the figure. For pairwise comparisons of microbial data, metatranscriptomic metrics were used for diversity and function abundance calculations. B) Microbial community diversity was significantly structured by stream order as depicted by non-metric multidimensional scaling of genome resolved metagenomic Bray-Curtis distances

(left, beta-diversity) and Pearson correlations of richness to stream order (right, alpha-diversity) with points colored by stream order. C) Sparse Partial Least Squares (sPLS) regressions show significant function (top) and MAG level (bottom) expression predictions of watershed maximum temperature, with key variables (Variable Importance Projection >1) denoted in bar graphs below. D) Non-metric multidimensional scaling of genome resolved metagenomic Bray-Curtis distances shows clustering of microbial communities by ecoregion (classified by Omernik II), with sampling location depicted on map above (mrpp, p<0.001). E) Random forest analysis showed that land use metrics could predict expressed gene richness with 72% accuracy, with order of variable importance shown as a bar graph. Abbreviations: NPOC, Non-Purgable Organic Carbon; DNRA, Dissimilatory Nitrite Reduction to Ammonia; WWTP Density, Waste Water Treatment Plant Density; NPP, Net Primary Production.

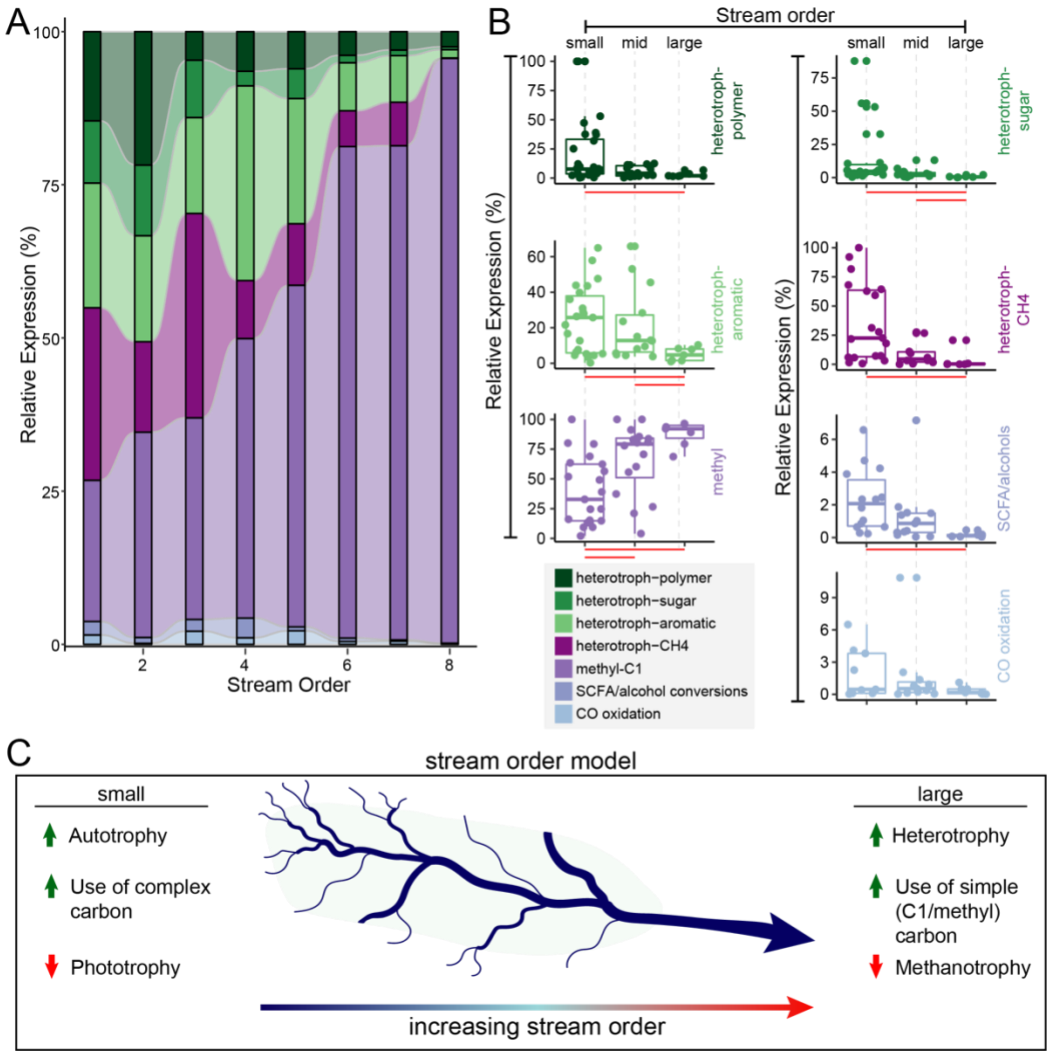
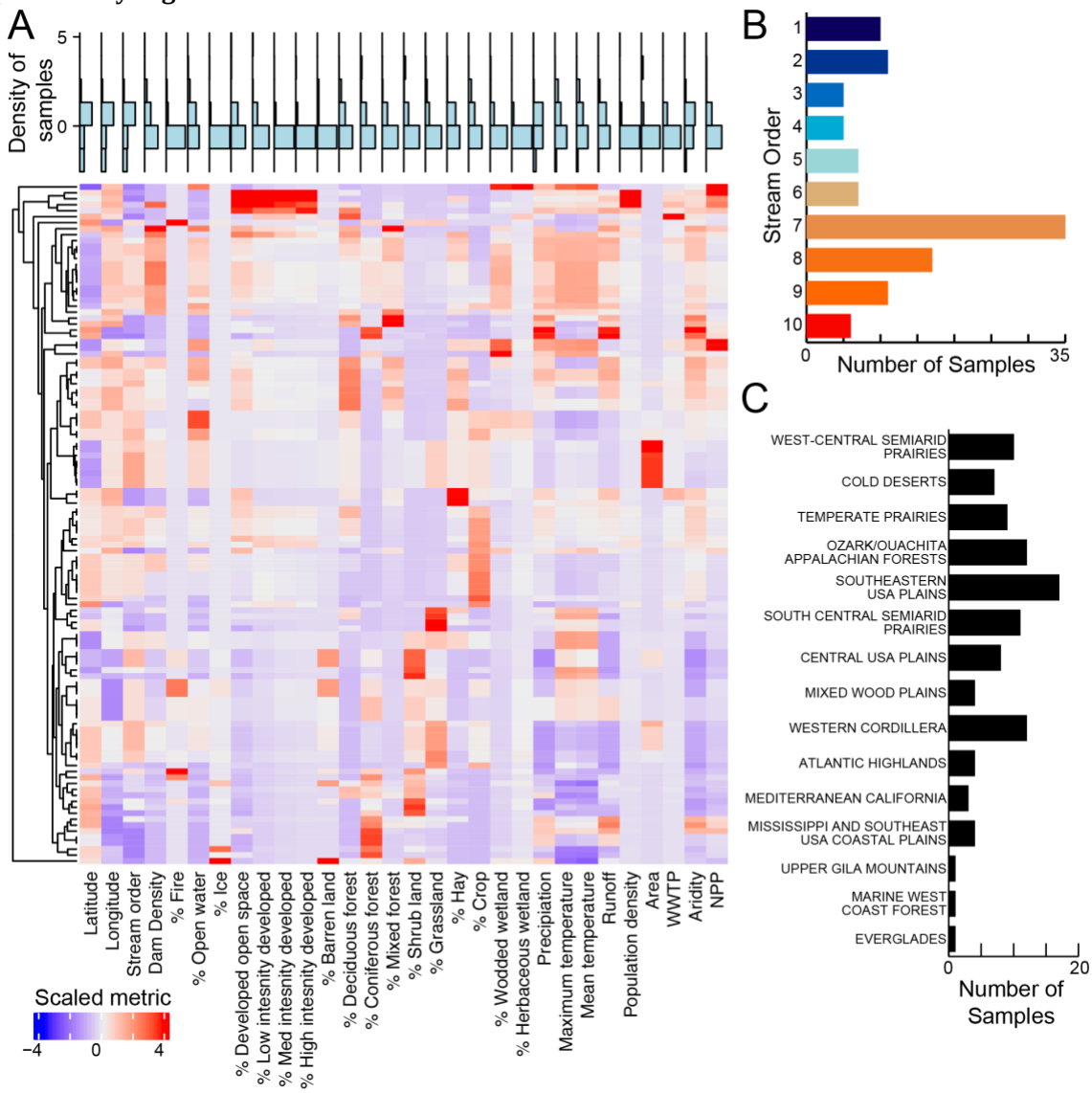


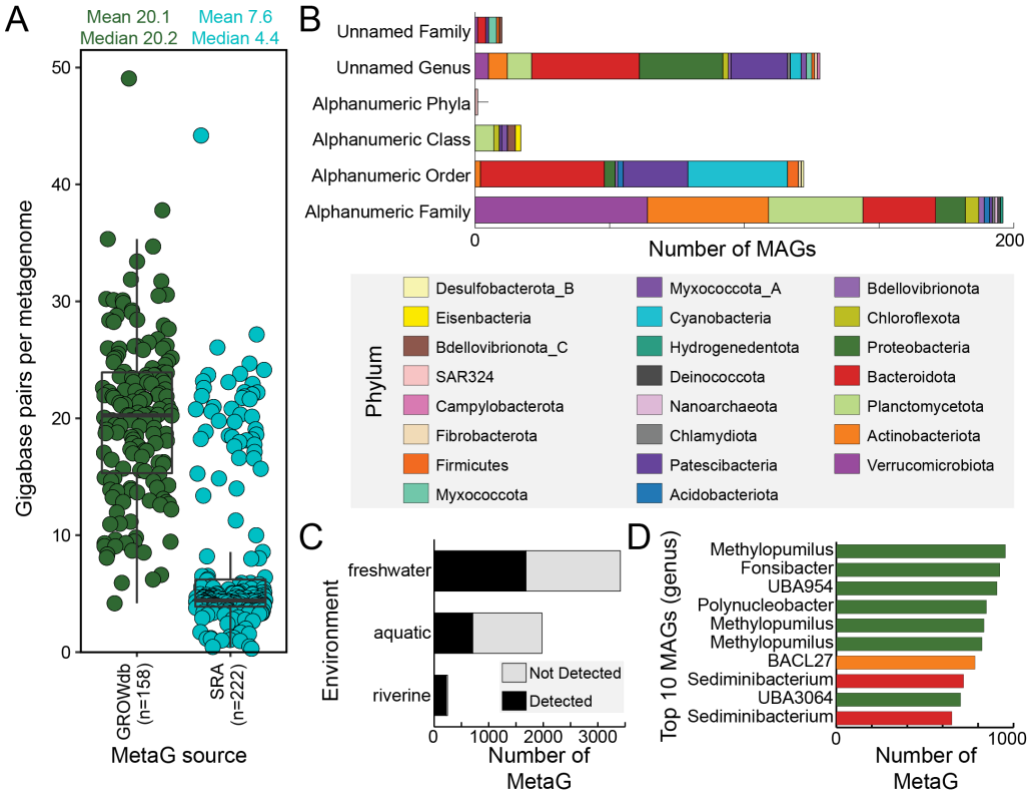
Figure 5: Microbial lifestyle and carbon utilization are structured along a stream order gradient. A) Alluvial plot shows the relative expression of microbial lifestyles (defined in methods) across stream order gradient. B) One dimensional scatterplots correspond to data in A, with each point representing a single sample and streams grouped by small (1-3), mid (4-6), and large (7-8) orders. Significant differences in expression (p<0.01) between small, mid, and large order streams are denoted by horizontal

red bars below each plot. C) The stream order model highlights changes in microbial expression from small (left) to large (right) order streams.

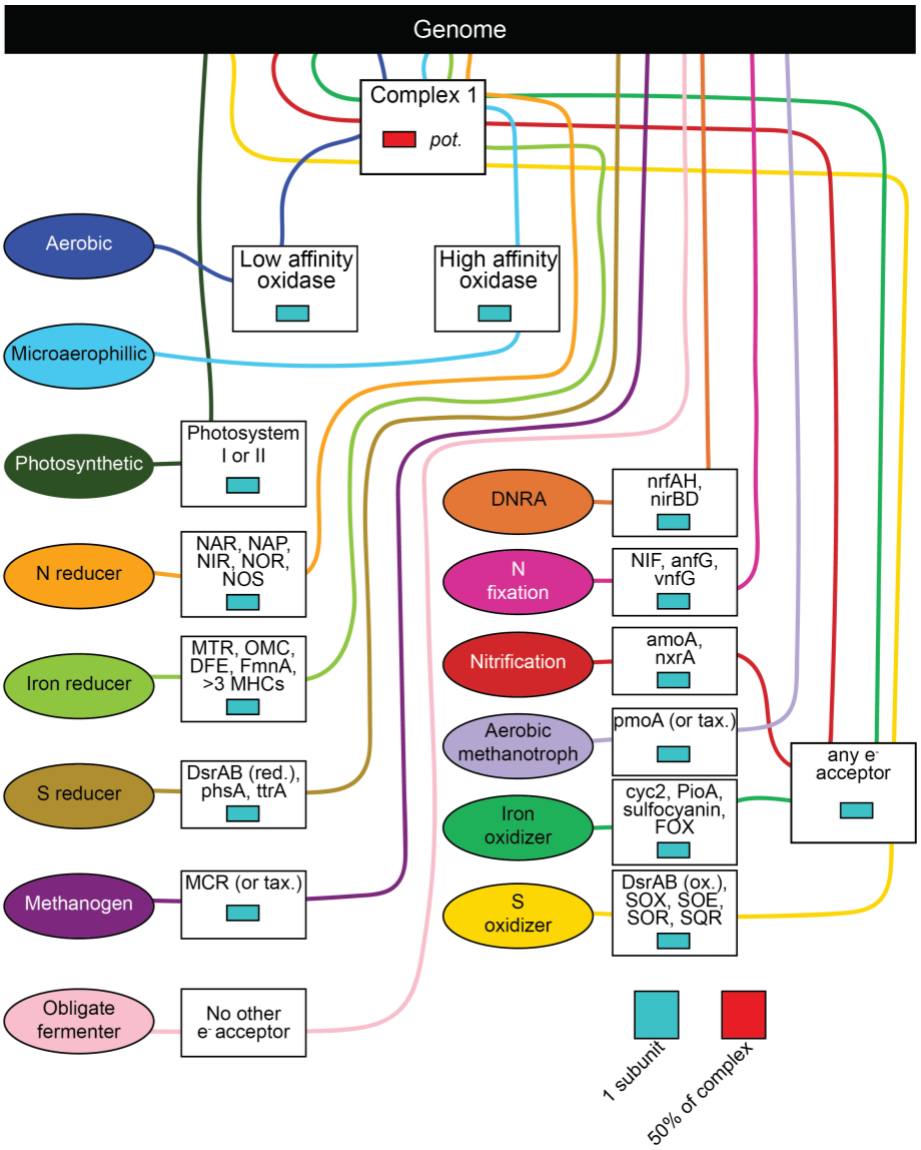
Supplementary Figures



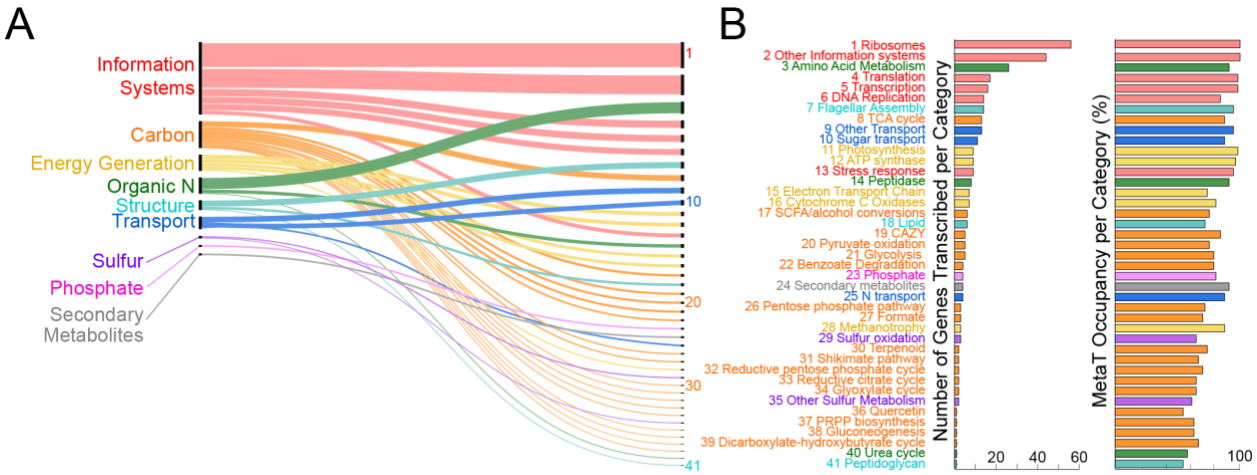
Supplementary Fig. 1: Distribution of river characteristics sampled in GROWdb. A) Heatmap of geospatial parameters sampled in GROWdb, where columns are environmental variables and rows are corresponding samples within GROWdb. Each variable has been scaled by subtracting the vector mean for each variable and dividing by its standard deviation. Blue histogram plots above highlight the distribution of samples for each variable, with high values at the top of the plot. Histogram plots of key variables used throughout the main text including stream order (B) and ecoregion (C) are also shown.



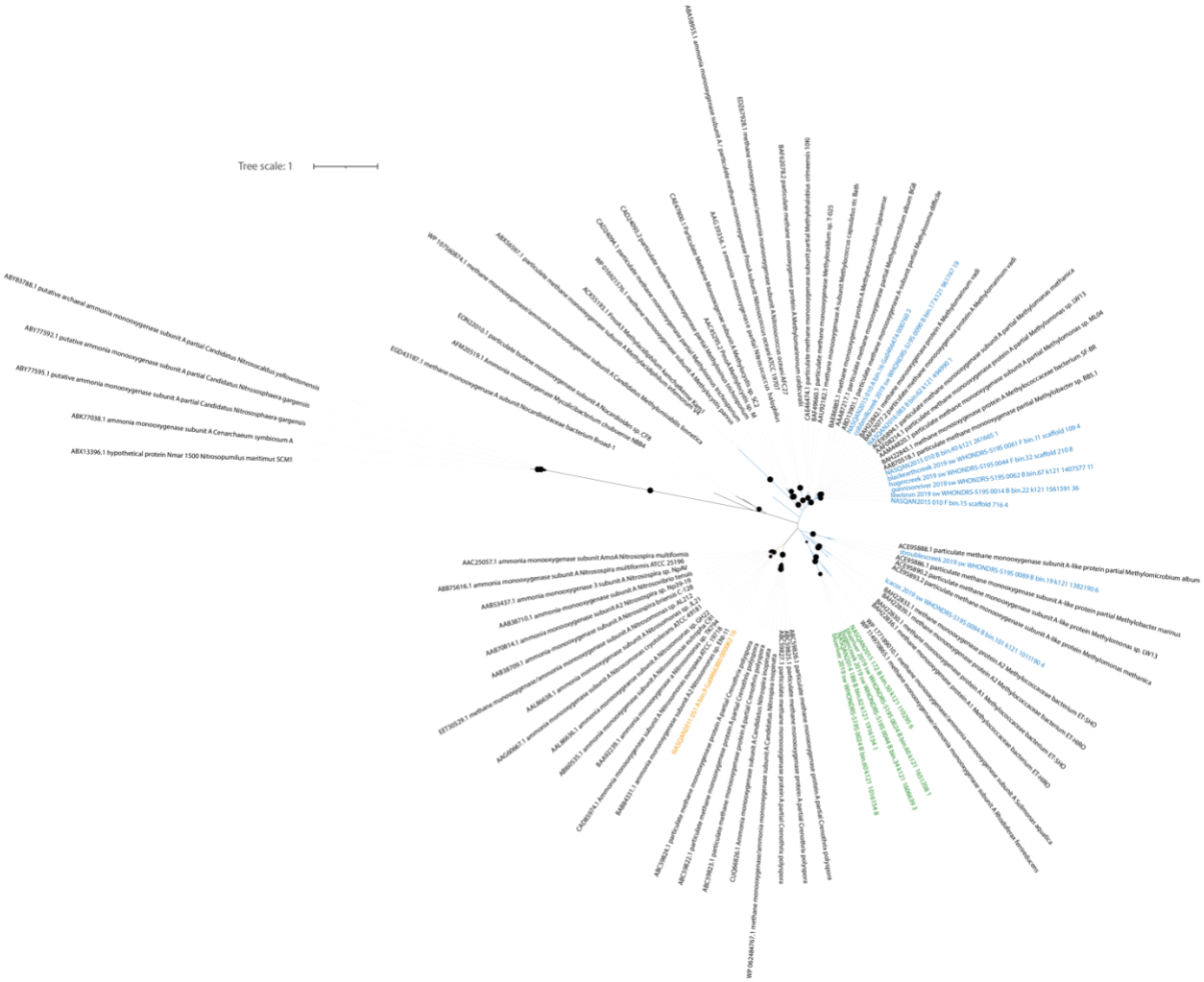
Supplementary Fig. 2: GROWdb comparison to other omics data sources. A) Sequencing depth comparison of GROWdb metagenomes (n=158) to SRA metagenomes classified as riverine (n=222) shows 3x increase in average sequencing depth for GROWdb. Each point represents a single metagenome, with mean and median values listed at the top of the graph. B) Stacked bar chart shows novelty of GROWdb MAGs when compared to GTDB (r207), where a MAG is placed at the highest level of novelty. Bars are colored by Phylum. C) Stacked bar chart shows the proportion of SRA metagenomes that GROWdb MAGs are detected (black) or not detected (grey) by environment category. D) The top ten MAGs most frequently detected across river surface water related SRA metagenomes are displayed at the genus level on the bar chart, with colors denoting phyla (key above).

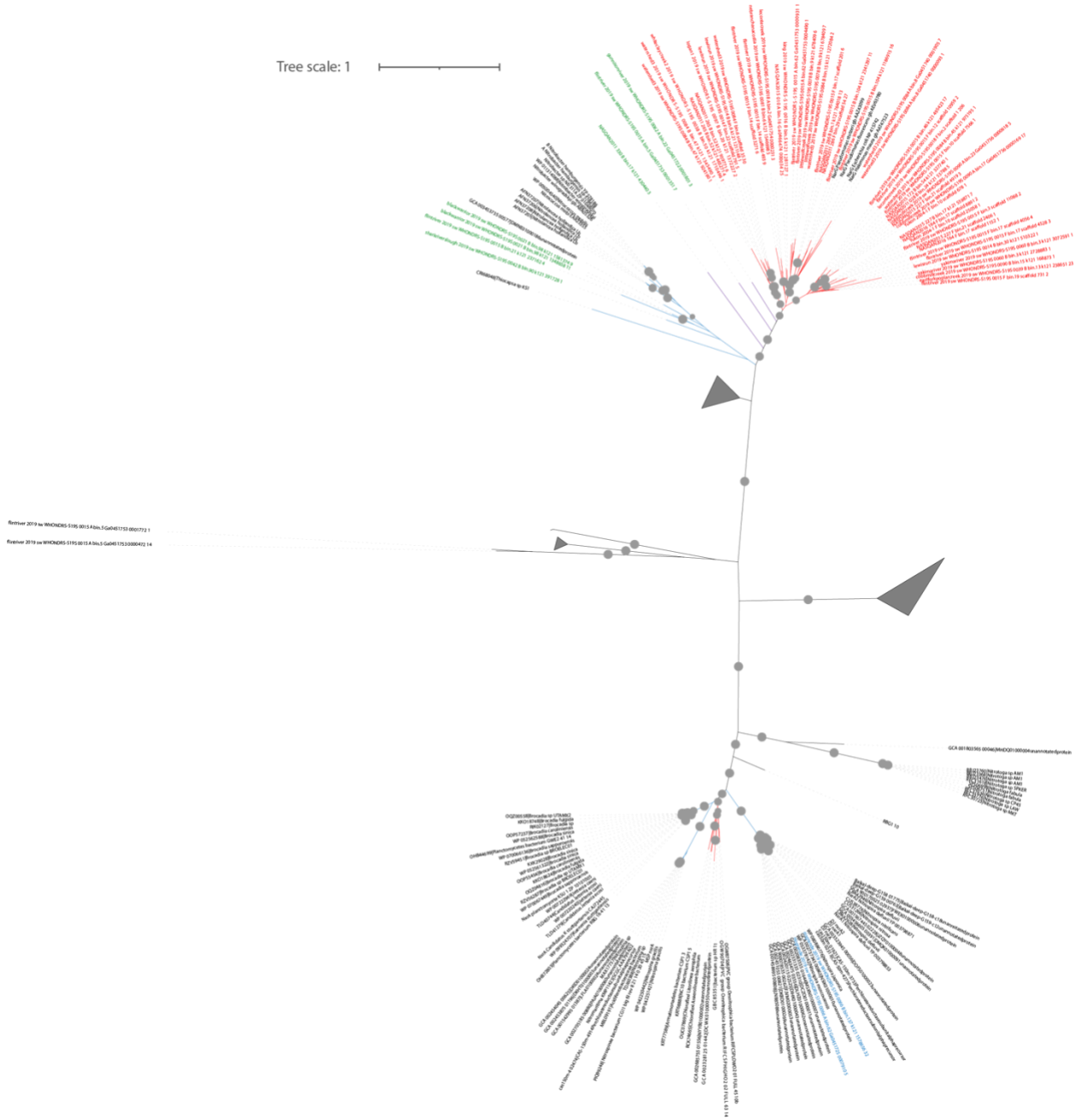


Supplementary Fig. 3: Metabolic trait assignment ruleset. Each trait is defined by a set of genes and the percent of genes required for that particular function. Lines flow from the genome (top black box) to traits (ovals), passing through boxes of gene requirements to be consider TRUE for that particular trait.

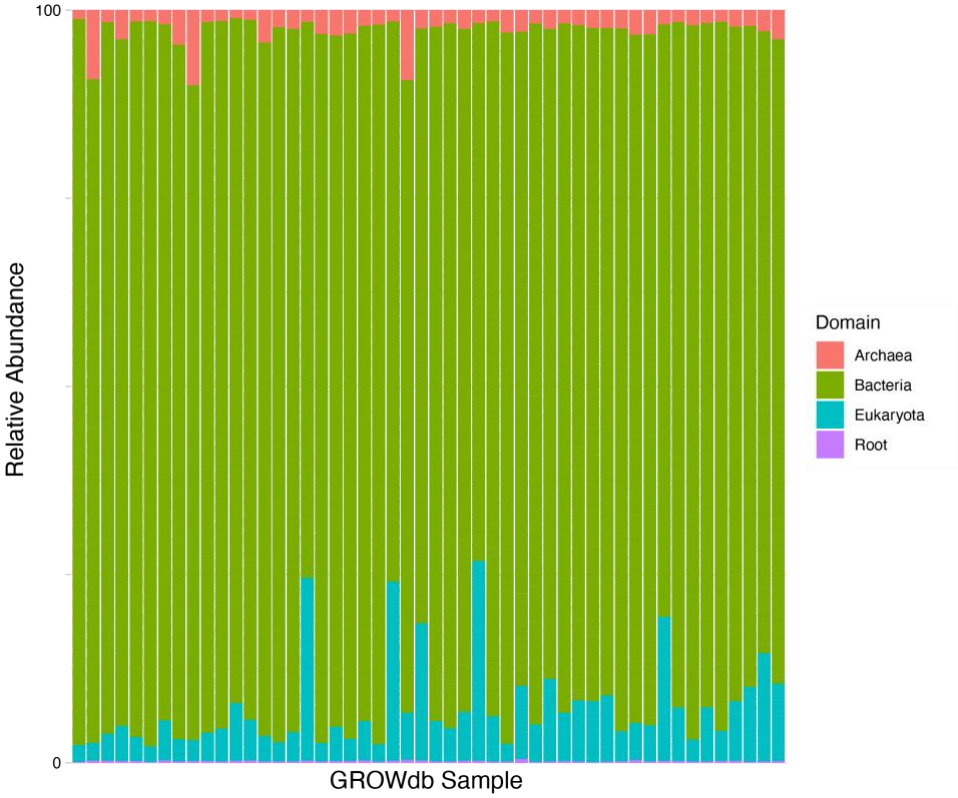


Supplementary Fig. 4: Gene level expression across rivers. Genes detected in more than 50% of metatranscriptomes, with gene functions (n=365) grouped by broad categories (n=9, A) and refined to subcategories (n=41, B). Thickness of lines and line order in A show the number of functions within a particular category (right) and subcategory (left). A and B are linked by subcategory number (1-41). For each of the 41 subcategories, the number of genes and occupancy defined as the percentage of samples detected across metatranscriptomes is shown by bar charts. Hypothetical and genes with unknown annotations are not shown, albeit 21 genes with these annotations were considered core or expressed in all metatranscriptomes.

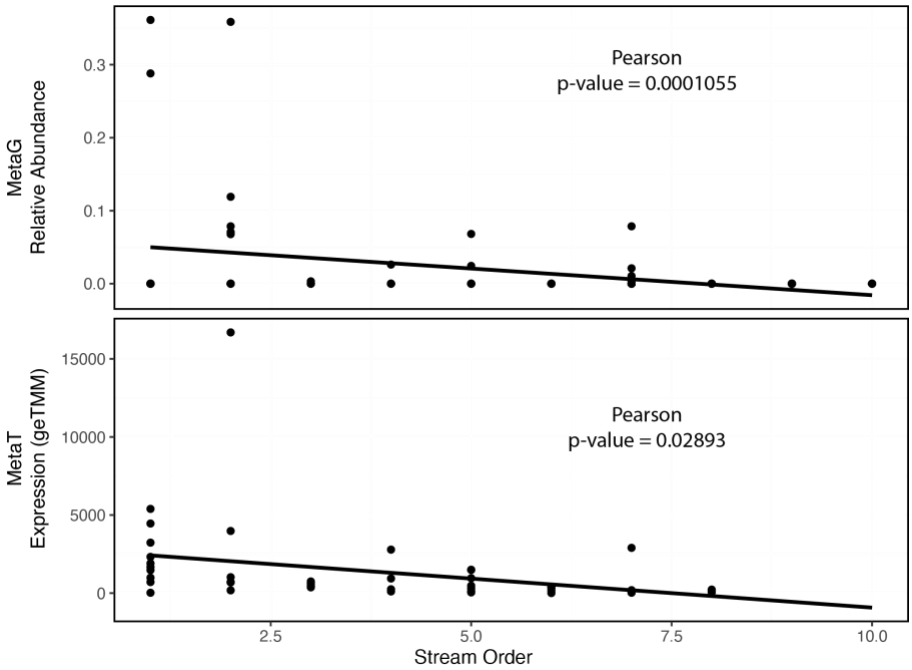

Tree scale: 1
ABY81788.1 putative archaeal ammonia monooxygenase subunit A partial Candidatus Nitrosocaldus yellowstonensis
ABY77592.1 putative ammonia monooxygenase subunit A partial Candidatus Nitrosophaera gargensis
ABY77595.1 putative ammonia monooxygenase subunit A partial Candidatus Nitrosophaera gargensis
ABK77038.1 ammonia monooxygenase subunit A Cenarchaeum symbiosum A
ABX13396.1 hypothetical protein Nmar 1500 Nitrosopumilus maritimus SCM1
AAC25057.1 ammonia monooxygenase subunit AmoA Nitrosospira multiformis
ABB75616.1 ammonia monooxygenase subunit A Nitrosospira multiformis ATCC 25196
AABS3437.1 ammonia monooxygenase 3 subunit A Nitrosomonas sp. Nm84
AAB38710.1 ammonia monooxygenase subunit A Nitrosomonas sp. Nm33
AAB38714.1 ammonia monooxygenase subunit A Nitrosomonas sp. Nm51
AAB38709.1 ammonia monooxygenase subunit A Nitrosomonas sp. Nm90
AAL86508.1 ammonia monooxygenase subunit A Nitrosomonas cryotolerans ATCC 49181
AAL86506.1 ammonia monooxygenase subunit A Nitrosomonas sp. AL212
AAL86504.1 ammonia monooxygenase subunit A Nitrosomonas sp. Is79A3
AAB60066.1 ammonia monooxygenase subunit A Nitrosomonas sp. Is79A7
AAB60105.1 ammonia monooxygenase subunit A Nitrosomonas sp. Is79A11
BAA82238.1 ammonia monooxygenase subunit A Nitrosomonas nitrosa strain Nm10
AABB3331.1 ammonia monooxygenase subunit A Nitrosomonas nitrosa strain Nm11
CAD86974.1 ammonia monooxygenase subunit A Nitrosomonas nitrosa strain Nm14
NASQMOS 2019 sw AAC-Scaffold00000 16
EET30529.1 methane monooxygenase/ammonia monooxygenase subunit A Candidatus Nitrosocaldus islandicus
ACE95888.1 particulate methane monooxygenase subunit A-like protein partial Methylomicrobium album
ACE95886.1 particulate methane monooxygenase subunit A-like protein partial Methylobacter marinus
ACE95895.1 particulate methane monooxygenase subunit A-like protein Methylomonas sp. LW13
ACE95892.1 particulate methane monooxygenase subunit A-like protein Methylomonas methanica
Kuenz 2019 sw WHONDRS-5195 0004 B bin.101 k121 1071190 4
BAA02930.1 methane monooxygenase protein A2 Methylococcaceae bacterium (T-SHO)
BAA02929.1 methane monooxygenase protein A1 Methylococcaceae bacterium (T-SHO)
BAA02928.1 methane monooxygenase protein A2 Methylococcaceae bacterium (T-HIRO)
BAA02927.1 methane monooxygenase protein A1 Methylococcaceae bacterium (T-HIRO)
ABC38014.1 particulate methane monooxygenase protein A Methylococcaceae bacterium
ABC38022.1 particulate methane monooxygenase protein A partial Methylomonas sp. LW13
ABC38023.1 particulate methane monooxygenase protein A partial Methylomonas methanica
CIO06806.1 ammonia monooxygenase/methane monooxygenase subunit A Candidatus Methylacidiphilum infernorum
WP 052484767.1 methane monooxygenase/ammonia monooxygenase subunit A partial Candidatus Methylacidiphilum fumariolicum
ABC38024.1 particulate methane monooxygenase protein A partial Methylomicrobium album
ABC38025.1 particulate methane monooxygenase protein A partial Methylomicrobium alcaliphilum
ABC38026.1 particulate methane monooxygenase protein A partial Methylomicrobium agile
ABC38027.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZR
ABC38028.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZM
ABC38029.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZK
ABC38030.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZL
ABC38031.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZN
ABC38032.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZO
ABC38033.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZP
ABC38034.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZQ
ABC38035.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZR
ABC38036.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZS
ABC38037.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZT
ABC38038.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZU
ABC38039.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZV
ABC38040.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZW
ABC38041.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZX
ABC38042.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZY
ABC38043.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ZZ
ABC38044.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AA
ABC38045.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AB
ABC38046.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AC
ABC38047.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AD
ABC38048.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AE
ABC38049.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AF
ABC38050.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AG
ABC38051.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AH
ABC38052.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AI
ABC38053.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AJ
ABC38054.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AK
ABC38055.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AL
ABC38056.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AM
ABC38057.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AN
ABC38058.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AO
ABC38059.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AP
ABC38060.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AQ
ABC38061.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AR
ABC38062.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AS
ABC38063.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AT
ABC38064.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AU
ABC38065.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AV
ABC38066.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AW
ABC38067.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AX
ABC38068.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AY
ABC38069.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. AZ
ABC38070.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BA
ABC38071.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BB
ABC38072.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BC
ABC38073.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BD
ABC38074.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BE
ABC38075.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BF
ABC38076.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BG
ABC38077.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BH
ABC38078.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BI
ABC38079.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BJ
ABC38080.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BK
ABC38081.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BL
ABC38082.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BM
ABC38083.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BN
ABC38084.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BO
ABC38085.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BP
ABC38086.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BQ
ABC38087.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BR
ABC38088.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BS
ABC38089.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BT
ABC38090.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BU
ABC38091.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BV
ABC38092.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BW
ABC38093.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BX
ABC38094.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BY
ABC38095.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. BZ
ABC38096.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CA
ABC38097.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CB
ABC38098.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CC
ABC38099.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CD
ABC38100.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CE
ABC38101.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CF
ABC38102.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CG
ABC38103.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CH
ABC38104.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CI
ABC38105.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CJ
ABC38106.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CK
ABC38107.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CL
ABC38108.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CM
ABC38109.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CN
ABC38110.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CO
ABC38111.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CP
ABC38112.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CQ
ABC38113.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CR
ABC38114.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CS
ABC38115.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CT
ABC38116.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CU
ABC38117.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CV
ABC38118.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CW
ABC38119.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CX
ABC38120.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CY
ABC38121.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. CZ
ABC38122.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DA
ABC38123.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DB
ABC38124.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DC
ABC38125.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DD
ABC38126.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DE
ABC38127.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DF
ABC38128.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DG
ABC38129.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DH
ABC38130.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DI
ABC38131.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DJ
ABC38132.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DK
ABC38133.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DL
ABC38134.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DM
ABC38135.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DN
ABC38136.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DO
ABC38137.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DP
ABC38138.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DQ
ABC38139.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DR
ABC38140.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DS
ABC38141.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DT
ABC38142.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DU
ABC38143.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DV
ABC38144.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DW
ABC38145.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DX
ABC38146.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DY
ABC38147.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. DZ
ABC38148.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EA
ABC38149.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EB
ABC38150.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EC
ABC38151.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ED
ABC38152.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EE
ABC38153.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EF
ABC38154.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EG
ABC38155.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EH
ABC38156.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EI
ABC38157.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EJ
ABC38158.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EK
ABC38159.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EL
ABC38160.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EM
ABC38161.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EN
ABC38162.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EO
ABC38163.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EP
ABC38164.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EQ
ABC38165.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ER
ABC38166.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ES
ABC38167.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ET
ABC38168.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EU
ABC38169.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EV
ABC38170.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EW
ABC38171.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EX
ABC38172.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EY
ABC38173.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. EZ
ABC38174.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FA
ABC38175.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FB
ABC38176.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FC
ABC38177.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FD
ABC38178.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FE
ABC38179.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FF
ABC38180.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FG
ABC38181.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FH
ABC38182.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FI
ABC38183.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FJ
ABC38184.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FK
ABC38185.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FL
ABC38186.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FM
ABC38187.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FN
ABC38188.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FO
ABC38189.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FP
ABC38190.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FQ
ABC38191.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FR
ABC38192.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FS
ABC38193.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FT
ABC38194.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FU
ABC38195.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FV
ABC38196.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FW
ABC38197.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FX
ABC38198.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FY
ABC38199.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. FZ
ABC38200.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GA
ABC38201.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GB
ABC38202.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GC
ABC38203.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GD
ABC38204.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GE
ABC38205.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GF
ABC38206.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GG
ABC38207.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GH
ABC38208.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GI
ABC38209.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GJ
ABC38210.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GK
ABC38211.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GL
ABC38212.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GM
ABC38213.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GN
ABC38214.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GO
ABC38215.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GP
ABC38216.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GQ
ABC38217.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GR
ABC38218.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GS
ABC38219.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GT
ABC38220.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GU
ABC38221.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GV
ABC38222.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GW
ABC38223.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GX
ABC38224.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GY
ABC38225.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. GZ
ABC38226.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HA
ABC38227.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HB
ABC38228.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HC
ABC38229.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HD
ABC38230.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HE
ABC38231.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HF
ABC38232.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HG
ABC38233.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HH
ABC38234.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HI
ABC38235.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HJ
ABC38236.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HK
ABC38237.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HL
ABC38238.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HM
ABC38239.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HN
ABC38240.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HO
ABC38241.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HP
ABC38242.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HQ
ABC38243.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HR
ABC38244.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HS
ABC38245.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HT
ABC38246.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HU
ABC38247.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HV
ABC38248.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HW
ABC38249.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HX
ABC38250.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HY
ABC38251.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. HZ
ABC38252.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IA
ABC38253.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IB
ABC38254.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IC
ABC38255.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ID
ABC38256.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IE
ABC38257.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IF
ABC38258.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IG
ABC38259.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IH
ABC38260.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. II
ABC38261.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IJ
ABC38262.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IK
ABC38263.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IL
ABC38264.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IM
ABC38265.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IN
ABC38266.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IO
ABC38267.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IP
ABC38268.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IQ
ABC38269.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IR
ABC38270.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IS
ABC38271.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IT
ABC38272.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IU
ABC38273.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IV
ABC38274.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IW
ABC38275.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IX
ABC38276.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IY
ABC38277.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. IZ
ABC38278.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JA
ABC38279.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JB
ABC38280.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JC
ABC38281.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JD
ABC38282.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JE
ABC38283.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JF
ABC38284.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JG
ABC38285.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JH
ABC38286.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JI
ABC38287.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JJ
ABC38288.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JK
ABC38289.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JL
ABC38290.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JM
ABC38291.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JN
ABC38292.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JO
ABC38293.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JP
ABC38294.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JQ
ABC38295.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JR
ABC38296.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JS
ABC38297.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JT
ABC38298.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JU
ABC38299.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JV
ABC38300.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JW
ABC38301.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JX
ABC38302.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JY
ABC38303.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. JZ
ABC38304.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KA
ABC38305.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KB
ABC38306.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KC
ABC38307.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KD
ABC38308.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KE
ABC38309.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KF
ABC38310.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KG
ABC38311.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KH
ABC38312.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KI
ABC38313.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KJ
ABC38314.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KK
ABC38315.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KL
ABC38316.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KM
ABC38317.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KN
ABC38318.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KO
ABC38319.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KP
ABC38320.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KQ
ABC38321.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KR
ABC38322.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KS
ABC38323.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KT
ABC38324.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KU
ABC38325.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KV
ABC38326.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KW
ABC38327.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KX
ABC38328.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KY
ABC38329.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. KZ
ABC38330.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LA
ABC38331.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LB
ABC38332.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LC
ABC38333.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LD
ABC38334.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LE
ABC38335.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LF
ABC38336.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LG
ABC38337.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LH
ABC38338.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LI
ABC38339.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LJ
ABC38340.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LK
ABC38341.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LL
ABC38342.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LM
ABC38343.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LN
ABC38344.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LO
ABC38345.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LP
ABC38346.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LQ
ABC38347.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LR
ABC38348.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LS
ABC38349.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LT
ABC38350.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LU
ABC38351.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LV
ABC38352.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LW
ABC38353.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LX
ABC38354.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LY
ABC38355.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. LZ
ABC38356.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MA
ABC38357.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MB
ABC38358.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MC
ABC38359.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MD
ABC38360.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ME
ABC38361.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MF
ABC38362.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MG
ABC38363.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MH
ABC38364.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MI
ABC38365.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MJ
ABC38366.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MK
ABC38367.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ML
ABC38368.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MM
ABC38369.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MN
ABC38370.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MO
ABC38371.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MP
ABC38372.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MQ
ABC38373.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MR
ABC38374.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MS
ABC38375.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MT
ABC38376.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MU
ABC38377.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MV
ABC38378.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MW
ABC38379.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MX
ABC38380.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MY
ABC38381.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. MZ
ABC38382.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NA
ABC38383.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NB
ABC38384.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NC
ABC38385.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. ND
ABC38386.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NE
ABC38387.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NF
ABC38388.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NG
ABC38389.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NH
ABC38390.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NI
ABC38391.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NJ
ABC38392.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NK
ABC38393.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NL
ABC38394.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NM
ABC38395.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NN
ABC38396.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NO
ABC38397.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NP
ABC38398.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NQ
ABC38399.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NR
ABC38400.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NS
ABC38401.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NT
ABC38402.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NU
ABC38403.1 particulate methane monooxygenase protein A partial Methylomicrobium sp. NV
ABC38404.1



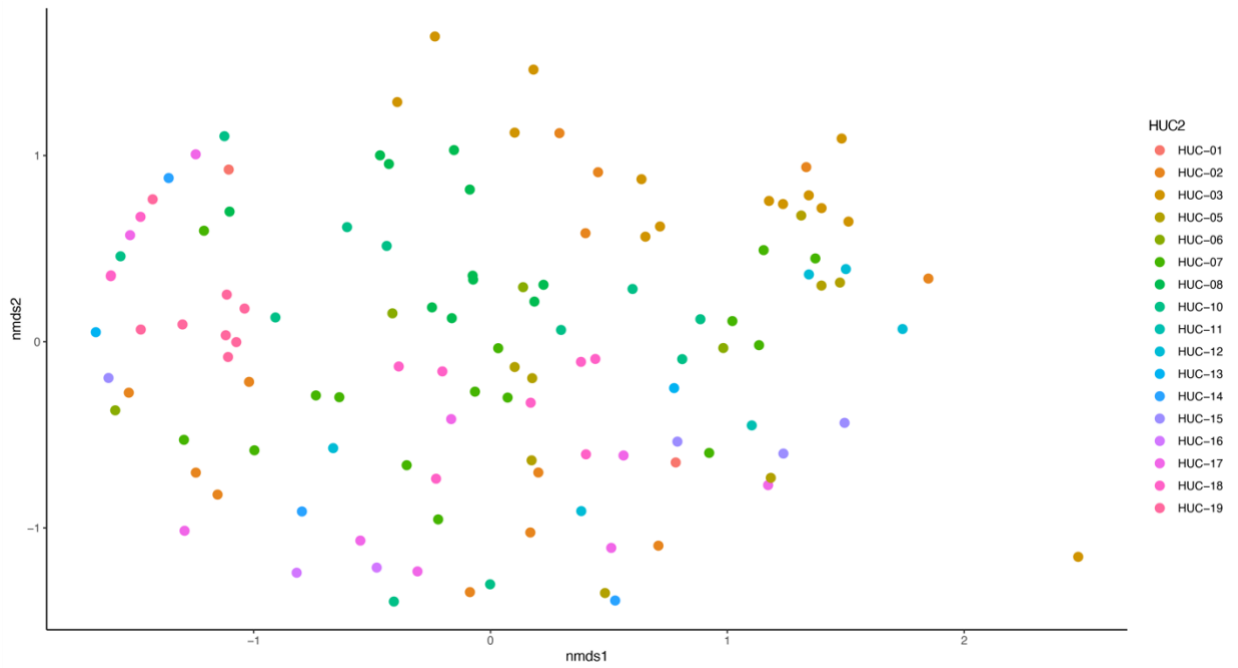
Supplementary Fig. 6: Phylogenetic analysis of nitrite oxidoreductase (nxr) and nitrate reductase (nar) is shown. GROWdb sequences colored by putative substrate, with nitrate (red), nitrite (blue), and unknown (green), while reference sequences are shown in black text. Bootstraps >80 are shown with closed grey circles at nodes.



Supplementary Fig. 7: Stacked bar chart shows the singleM profiles of GROWdb metagenomic reads, with bars colored by domain. By domain, the most reads constituted the Bacteria (mean=90%), followed by Eukaryota (mean=7.1%), Archaea (mean=2.7%), and Unknown (mean=0.2%).



Supplementary Fig. 8: Correlations of Patescibacteria relative abundance (metagenomics, top) and expression (metatranscriptomics, bottom) with stream order.



Supplementary Fig. 9: Microbial community diversity was significantly structured by hydrologic unit 2 (HUC2) reagion as depicted by non-metric multidimensional scaling of genome resolved metagenomic Bray-Curtis distances.