A functional microbiome catalog crowdsourced from North American rivers

Mikayla A. Borton[1]*, Bridget B. McGivern[1], Kathryn R. Willi[2], Ben J. Woodcroft[3], Annika C. Mosier[4], Derick M. Singleton[4], Ted Bambakidis[5], Aaron Pelly[6], Filipe Liu[7], Janaka N. Edirisinghe[7], José P. Faria[7], Ikaia Leleiwi[1], Rebecca A. Daly[1], Amy E. Goldman[8], Michael J. Wilkins[1], Ed K. Hall[2], Christa Pennacchio[9], Simon Roux[9,10], Emiley A. Eloe-Fadrosh[9,10], Stephen P. Good[11], Matthew B. Sullivan[12], Christopher S. Henry[7], Elisha M. Wood-Charlson[10], Matthew R.V. Ross[2], Christopher S. Miller[4], Byron C. Crump[11], James C. Stegen[6,13], 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] Terrestrial and Aquatic Integration, Pacific Northwest National Laboratory, Richland, WA, USA
[7] Mathematics and Computer Science Division, Argonne National Laboratory, Argonne, IL, USA
[8] Earth Systems Science Division, Pacific Northwest National Laboratory, Richland, WA, USA
[9] DOE Joint Genome Institute, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[10] Environmental Genomics and Systems Biology Division, Lawrence Berkeley National Laboratory, Berkeley, CA, USA
[11] College of Earth, Ocean, and Atmospheric Sciences, Oregon State University, Corvallis, OR, USA
[12] Department of Microbiology, The Ohio State University, Columbus, OH, USA
[13] School of the Environment, Washington State University, Pullman, WA, USA

Correspondence* 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 processes 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, sequencing, and genome-resolved analyses 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. Specifically, GROWdb encompasses 1,469 microbial species from 27 phyla, including novel lineages from 10 families and 128 genera, and defines the core river microbiome for the first time at genome level. GROWdb analyses coupled to extensive geospatial information revealed local and regional drivers of microbial community structuring, while also presenting a myriad of foundational hypotheses about ecosystem function. Building upon the previously conceived River Continuum

Concept[1], we layer on microbial functional trait expression, which suggests the structure and function of river microbiomes is predictable. We make GROWdb available through various collaborative cyberinfrastructures[2,3] so that it can be widely accessed across disciplines for watershed predictive modeling and microbiome-based management practices.

Main

Earth's surface is dominated by water, much of it the oceans, which is known to buffer against anthropogenic climate change via microbes dictating the fate of ocean-absorbed carbon[4]. While the oceans and their microbes have been extensively studied globally by large scientific consortia (e.g. *Tara* Oceans Consortium[5]), other elements of Earth's water system, such as rivers, are relatively understudied. This is problematic as rivers (i) offer an important nexus of nutrient transport across terrestrial and aquatic interfaces[6], (ii) are hotspots for biogeochemical processes that significantly contribute to global terrestrial carbon and nitrogen budgets ultimately impacting global greenhouse gas (GHG) emissions, eutrophication, and acidification[6–8], and (iii) have immediate societal impacts on sustainable energy, agriculture, environmental health, and human health[9–11]. Microbial metabolisms dictate river ecosystem functioning with major influence on carbon (C) respiration and sequestration, nitrogen (N) cycling and uptake, food webs, and pollutants (e.g., Hg, As)[12–16]. 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 remain under sampled due to methodological constraints and sampling scope. For example, a majority of river microbiome studies rely only on 16S rRNA gene analysis (**Extended Data File 1**). While these single gene studies have advanced understanding of riverine microbial community diversity and membership[17–20], they lack information on poorly characterized lineages and are limited in their capacity to functionally link microorganisms to biogeochemical processes. Albeit less, there are several studies with metagenomics that provide functional attributes of river microbiomes, but these rely heavily on existing databases rather than *de novo* assembly and metagenome assembled genomes (**Extended Data File 1**), masking the

contributions of novel members of the microbiome. Less than five studies employ genome-resolved expression methods hindering estimates of the metabolic processes that are active in river systems (**Extended Data File 1**). Finally, in terms of sampling, most studies focus on a single site or stream network (**Extended Data File 1**), leaving the generalizability of microbiome rules 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 a crowd-sourced, distributed sampling effort to increase and standardize river microbiome sampling globally, compiling these sequencing results into a large-scale Genome Resolved Open Watersheds database (GROWdb). An emphasis of GROWdb is a publicly available and ever-expanding microbial genome database that has accessible content. 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. GROWdb is based on a crowd-sourced, network-of-networks approach[21] to move beyond a small collection of well-studied rivers, towards a spatially distributed, global network of systematic observations.

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

To establish the GROWdb, over 100 teams were crowd-sourced to broadly sample 163 sites across United States rivers and develop ~3.8 terabases of metagenomic and metatranscriptomic sequencing data to go with extensive (up to 287) geochemical and geospatial measurements at each site (**Fig. 1A**, detailed per-sampling data available in **Fig. 1B** and **Extended Data File 1**). Through this, we aimed to capture community-level, genome-resolved microbiome variations in taxonomy, function, and gene expression in the context of geographic and environmental gradients across the United States. The effort resulted in surface water sampling that covered 90% of United States watersheds (n=21 as determined by hydrologic unit 2) (**Fig. 1C**) and spanned diverse ecoregions, stream orders, and watershed

sizes (**Extended Data Fig. 1**). Collectively, GROWdb integrates genomics, biogeochemistry, and a broad range of contextual environmental variables into a platform that enables a predictive framework of microbiomes and their biogeochemical contributions.

To ensure data accessibility we provide four access points for user engagement with GROWdb (**Fig. 1A**). First, all reads and MAGs are publicly hosted on National Center for Biotechnology (NCBI), enabling transferability to resources that pull and incorporate this content (e.g., Genome Taxonomy Database[22]). Datasets underlying GROWdb are freely available and searchable through the National Microbiome Data Collaborative (NMDC)[23] data portal, linking to other data types (e.g., metabolome[24]) to allow for broader synthesis where available. GROWdb MAGs are available as an annotated genomic collection in the freely accessible KBase[3] cyberinfrastructure. Here 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 distilled the taxonomic and functional insights from GROWdb in a web accessible format, called GROWdb Explorer, allowing the rapid profiling of taxonomic and functional distributions across the data set. GROWdb version 1 can be accessed across platforms (**Fig. 1A,** QR codes), making this microbiome content available in an expanding repository to incorporate and unify global river multi-omic data for the future.

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[25], thereby increasing the sensitivity for detecting the breadth of microbial functions encoded in these systems (**Extended Data 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, Extended Data File 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 sequencing reads were well represented by the genomic 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[18,26–28]. Beyond providing new genomic resources for these ecologically "known" taxa, the GROWdb MAGs provide new genomic resources for many less well-known taxa. A subset of our genomes represented newly sampled lineages, including 10 families and 128 genera across 16 phyla (**Extended Data Fig. 2**). Additionally, a large proportion of MAGs belonged to lineages defined only by alphanumeric names (e.g., Uncultured Bacterial and Archaeal genomes, UBA[29]) at the phylum (n=1), class (n=17), order (n=121), and family (n=196) levels (**Extended Data Fig. 2**). Highlighting the relevance of GROWdb, analysis of 266,764 public metagenome datasets in the Sequence Read Archive (SRA)[25] 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 taxonomic assignments). We verified the most prevalent phyla and genera in GROWdb had parallel representation in publicly available metagenomes (**Extended Data 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 harbor diverse communities across habitats acting as integrators across landscapes (**Fig. 2B**).

The comparison to publicly available metagenomes underscored the need for this river-based microbiome study, as freshwater related metagenomes were only half to a third as many as their soil and ocean counterparts respectively in the SRA[25]. 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 practically unusable. GROWdb ascribes to standardized protocols and metadata practices[2,30,31], 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**). Within these phyla, genera that were the most cosmopolitan (occupancy; number of samples) across samples, were also the most abundant members of these communities (**Fig. 3B**). This was especially true for MAGs affiliated with the genus *Planktophilia*, a well-known freshwater microorganism[32], that were present in 70% of the GROW metagenomes, and had the highest mean relative abundance across samples 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. 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, indicating that some microbial strains have widespread metabolic activity across rivers. Here we show how analyses of GROWdb allows us to constrain the thousands of microbial genomes to a most-wanted list of the most active microorganisms in river systems, identifying lineages and metabolic pathways that could represent diagnostic or metabolism targets needing accurate representation in biogeochemical models moving forward.

To understand the impacts of these core, transcriptionally 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 (**Extended Data Fig. 3, Extended Data File 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 oxygenic or anoxygenic photosystems or simple, low yield photorhodopsins. In fact, of the top 25 most active genera, more than 90% were capable of aerobic respiration or light driven metabolisms, with many encoding multiple light harvesting mechanisms (**Fig. 3C, Extended Data 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 (**Extended Data 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) (**Extended Data Fig. 6**). Together the encoding of both aerobic and anaerobic energy systems, and diverse electron donors across the many core, active taxa highlight the metabolic versatility harbored in river 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 (**Extended Data Fig. 7**). We identified one MAG within the bacterial *Nitrosomonas* genus that encoded 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[33].

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[16,34] is unlikely as these MAGs contained genes for nitrite oxidation but lacked genes for

ammonia oxidation. These two 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[16]. 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 that were transcriptionally active from the 48 total MAGs sampled in this phylum. These genomes all lacked the capacity for aerobic or anaerobic respiration and were inferred to be anoxic, obligate fermenters, consistent with prior genomic reports from this phylum that to date lacks any pure culture, characterized representatives[35,36]. Given surface waters are oxic, we verified the abundance patterns reported here were consistent with other river metagenome and amplicon-based studies[37,38], 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 exist as endosymbionts, or thrive in lower oxygen niches associated with suspended particles, biofilms or hyporheic environments where oxygen can be depleted during dissolved organic matter decomposition[39,40]. In support of the latter, we observed that relative abundance and expression of Patescibacteria significantly decreased with river size (**Extended Data Fig. 8**), suggesting these obligate fermenters were more active in shallow waters when there is greater exchange between water and the stream bed[41].

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[42]. Antibiotic production is also part of natural competition in these complex communities. We cataloged 1,587 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)[43]. Individual MAGs sometimes coded for ARGs from multiple gene families and targeting multiple drug classes.

Interestingly, most (n=1,219) 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[44,45], 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[44,45], 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, allowing us to contextualize the factors driving microbial biogeography at the continental scale. Previous studies have done this at local or regional levels using taxonomy alone[17,20,46], but to our knowledge these analyses have not incorporated functional gene-trait information nor been done at the continental scale. 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 used to identify the potential drivers of microbiome structure and expressed function (**Extended Data File 1**).

Of all the river site variables examined (**Fig. 4A**), stream order, a numerical ranking of the relative river size that spans small headwater streams (low order 1-3) to larger rivers like sections of Mississippi river (high order 8-12), was the most important controller of microbiome composition. River size was more important than latitudinal position or total carbon, which are often cited as controllers of microbiomes across other habitats[47,48]. Both metagenomes and metatranscriptomes were structured by stream order (**Fig. 4A and 4B**), providing evidence in favor of the River Continuum Concept[1], detailed below. After stream order, expressed microbial functional profiles were also influenced by watershed air temperature (both mean and maximum), area, and total runoff (**Fig. 4A**).

Given this relationship with local atmospheric conditions, we sought to understand which functional traits and microorganisms most contributed 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 air 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 suggest that climate influences on water temperature play a defining role in the niches of these microorganisms. These findings are consistent with reports from marine systems[49], hinting at 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[50], 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 (**Extended Data Fig. 9**). These findings support earlier work examining stream microbial succession, which showed that river microbial communities are inoculated from the landscape,

and this terrestrial influence continues to play an important role in downstream community assembly processes[19]. We note that the spatial structuring was not observed at the expressed functional level, 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[51,52].

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[53]. Particularly important is how different land management types may structure surface water microbiomes. Here we used a machine learning approach to associate river MAGs or gene content and expression 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 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[1]. 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 applicability of the RCC to microbial communities has been observed as decreased microbial 16S rRNA gene richness occurring across stream order gradients in Toolik[19], Danube[54], Mississippi[53], and Amazon[55] rivers. Here we show that the RCC also applies to

microbial community genomic and transcriptomic patterns at the continental scale (**Fig. 4B**). Importantly, this discovery provides 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 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[53]. To our knowledge this is the first report of genome-resolved metatranscriptomics across rivers and suggests that genome inferred transcriptional richness may be governed by a different set of environmental controls than gene presence at the continental scale.

One major control on biological diversity described by the RCC is variability in sunlight exposure. Lower order streams are often characterized by thick shore vegetation or overhanging trees that limit sunlight penetration and restrict phytoplankton and benthic microalgae primary production[1,56]. Consistent with this idea, we observed statistically significant increase in light-driven microbial metabolisms when moving from lower order streams to higher 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. Using microbial gene expression coupled to genome-resolved lifestyle information we estimated 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.

Another ecological control described by the RCC is a downstream decrease in the importance of terrestrial carbon inputs. We 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, transcript of genes targeting polymers, aromatics, and sugars are upregulated in low order streams, while methylotrophy gene transcripts, primarily from methanol oxidation, increased in higher order rivers. We consider methanol is likely autochthonous, derived from river phytoplankton biomass[57] or microbial metabolism of aromatic allochthonous plant litter[58,59]. 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 higher order rivers.

In summary, river systems were once thought of as passive pipes, transporting water from terrestrial to marine systems. As a result, it was 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. This study also supports the application of the RCC to microbial communities and provides the first evidence that landscape patterns in river microbiomes are grounded in mechanistic changes in genomic function. 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 adds a new view of how microbial metabolism changes across rivers.

Conclusion

Rivers are a critical part of the Earth System for their roles in dispersing water across terrestrial ecosystems. Changing climate impacts rivers via altered precipitation intensity, surface runoff, flooding, fires, sea level rise, and droughts, and all of these have direct impacts to human health, agriculture, energy production, and ecosystem resiliency[60]. Additionally, 2/3 of drinking water in the United States come

from surface river waters. Consequently, river management is expected to be one of the most politically charged topics in decades to come[61]. Microbes are master orchestrators of nutrient and energy flows that will likely dictate water quality under current and future water scenarios.

The Earth System is at a breaking point and microbiomes may be one of the levers to push back on climate chaos[4]. The field of environmental microbiology has undergone rapid transition from microscopy revealing abundant microbes to gene markers demonstrating the vast diversity of microbiota colonizing Earth's habitats[62]. Today's challenge is linking "who is there" to their functional and ecosystem roles and doing so at scale that allows elucidation of emergent properties from diverse and complex systems. Here we crowdsourced a large-scale river microbiome study using standardized sampling, processing, sequencing, and analysis to enable cross-site comparisons and modular augmentation. This resulted in a comprehensive resource that includes a collection of over two thousand curated metagenome-assembled genomes. We show this resource can be used to rapidly contextualize "who" and "what processes" microbes are modulating in river systems. This product and its many data access and synthesis sites reduces the computational barriers for expediting the translation of reads to functional content. Ultimately, GROWdb is a highly contextualized, genome-resolved resource spanning river sizes, locations, and land management regimes to promote microbiota derived insights across disciplinary boundaries.

GROWdb offers a genome-centric window into river microbiota and a FAIR-use cyberinfrastructure-powered platform for future researchers. GROWdb analyses presented here reveals core functionalities of river microbiomes and illuminates key metabolisms and microorganisms that are conserved and are transcriptionally active across the United States river systems. We envision that this genomic infrastructure will pave the way for future developments in water quality monitoring and identifying biomarkers indicative of land use or water quality changes. 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 two approaches to obtain samples from across United States rivers. One was a network-of networks[21] approach based on sampling efforts of the Worldwide Hydrobiogeochemistry Observation Network for Dynamic River Systems (WHONDRS) consortium[63], which is designed to facilitate the development of transferable scientific understanding and mutual benefit across stakeholders[30,31]. 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. The other approach was through a collaboration with the U.S. Geological Survey (USGS) National Water Quality Network (NWQN)[64]. The rest of the samples were collected under the NWQN campaign, which is a USGS long-term water-quality monitoring program to characterize consistent information on streamflow and water-quality conditions. Data are collected to assess the status and trends of water-quality conditions at large inland and coastal river sites, as well as in small streams indicative of urban, agricultural, and reference conditions[64].

Samples described in GROWdb version 1 were collected under the WHONDRS 2019 sampling campaign (**Extended Data File 1**) and 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. Geochemical data collected under the WHONDRS 2019 sampling campaign are available ESS-DIVE, with methods described[24]. For microbiome analyses, at 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 hours of collection. Surface water samples and filters were immediately frozen at −20 °C upon receiving for nucleic acid extraction, respectively. Methods of sample

collection used by the National Water Quality Network (NWQN) conform to the USGS National Field Manual for the Collection of Water-Quality Data[66], and DNA was collected on 0.22 μm Sterivex-GP filter (EMD Millipore). Here we provided kits integrated with USGS protocols for river sample processing with samples were preserved as described previously[67]. All samples were stored on ice and stored at –20°C until nucleic acid extraction.

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 (JGI) 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 kits and methods 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[68]. 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 in this time frame 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[69] and the National Hydrography Plus Version 2 (NHDPlus V2) Dataset using the *nhdplusTool*s package[70] in R. 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[71] 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 **Extended Data File 1** for a complete list of datasets included in our analysis. The complete R workflow for this geospatial analysis can be found on GitHub[72].

Metagenomic assembly, binning, and annotation

At the Joint Genome Institute, genomic DNA was prepared for metagenomic sequencing using Plate-based DNA library preparation on the PerkinElmer Sciclone NGS robotic liquid handling system. 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 accessible GROWdb pipelines released on GitHub[68]. Briefly, to maximize genome recovery three assemblies were performed on each set of fastq files and binned separately: (1) Read trimming with sickle (v1.33)[73], assembly with megahit (v1.2.9)[74], and binning with metabat2[75] (2.12.1) (2) Read trimming with sickle (v1.33)[73], random filtering to 25% of reads, assembly with idba-ud[76] (1.1.0), and binning with metabat2[75] (2.12.1) (3) Bins derived from the

JGI-IMG pipeline[77] were downloaded. All resulting bins were assessed for quality using checkM[78] (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[79] (v2.6.2) to obtain the dereplicated first version of the GROW database (n=2,093 MAGs). MAG taxonomy was assigned using GTDB-tk[22] (v2.1.1, r207) and annotated using DRAM (v1.4.4)[80].

To quantify MAG relative abundance across samples, trimmed metagenomic reads were mapped to the dereplicated MAG set using Bowtie2[81] 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 the mean coverage for each MAG, we used CoverM[82] (-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[82] 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[68].

Metatranscriptomic mapping and analysis

RNA was prepared for 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[81] (-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[83]. Counts were transformed to geTMM (gene length corrected trimmed mean of M-values) in R using edgeR package[84]. Genes and bins were considered if they were expressed in 10% of samples. Core calculations in **Fig. 3** had an additional requirement to express at least 20 genes.

Microbial metabolism trait and carbon usage classification

To classify microbial genes and genomes based on their carbon metabolism we curated the metabolism assignments made by DRAM[80] using rulesets to assign genomes to functional guilds (**Extended Data Fig. 3**). For example, genomes were classified by respiratory capacity based on the presence of >50% of the subunits required for Complex 1 of the electron transport chain and the presence at least one gene for an electron acceptor. As such, for a genome to be classified as a microaerophile, we required the genome to have more than 50% of complex 1 subunit and at least one subunit of a low affinity cytochrome oxidase. Likewise, if a genome did not have more than 50% of the subunits required for Complex 1 of the electron transport chain or the potential for any electron acceptor, it was classified as an obligate fermenter (**Extended Data 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 genes required for utilizing different carbon substrates in the genomes identified for heterotrophy. We assigned carbon gene expression into the following categories: polymer, sugar, aromatic compound, methanotrophy, methylotrophy, short chain fatty acid utilization, and carbon monoxide utilization using DRAM assigned rules. Carbon usage curation scripts are available on the GROWdb GitHub[68]. P/R ratios were defined by the ratio of expression of light driven energy metabolisms (aerobic photosynthesis, anaerobic

photosynthesis, and photorhodopsins) divided by aerobic respiration metabolisms (aerobic respiration and microaerophilic respiration).

Phylogenetic analyses were performed to refine the annotation of nitrogen related metabolism including genes annotated as respiratory nitrate reductase (*nar*), nitrite oxidoreductase (*nxr*), ammonia monooxygenase (*amo*), or methane monooxygenase (*pmo*) to improve the assignment the nitrogen cycling capabilities of GROW MAGs (**Extended Data File 3**). Specifically, Nxr/Nar and PmoA/AmoA amino acid reference sequences were downloaded[34,85,86] and this 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[87]. This resulted in two phylogenies, one for Nxr/Nar and one for Pmo/Amo, with this homology based approach used to refine the homology based gene annotations in the MAG database.

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)[43]. 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).

Statistical Analysis

All data analysis and visualization was done in R (v4.2.1) with the following packages: stats (v4.1.1), vegan (v2.6), ggplot2 (v.3.3.6), ComplexUpset (v2.8.0), tidyr (v1.2.0), dplyr (v1.0.9), corrplot (v0.92), pheatmap (v1.0.12), RColorBrewer (v1.1-3), pls (v2.8), edgeR (v3.16). Scripts for figure generation and data analysis are available on GitHub[68].

References

1. 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).
2. 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).
3. Arkin, A. P. *et al.* KBase: The United States Department of Energy Systems Biology Knowledgebase. *Nat. Biotechnol.* **36**, 566–569 (2018).
4. Cavicchioli, R. *et al.* Scientists' warning to humanity: microorganisms and climate change. *Nat. Rev. Microbiol.* **17**, 569–586 (2019).

5. Sunagawa, S. *et al.* Tara Oceans: towards global ocean ecosystems biology. *Nat. Rev. Microbiol.* **18**, 428–445 (2020).
6. Battin, T. J. *et al.* River ecosystem metabolism and carbon biogeochemistry in a changing world. *Nature* **613**, 449–459 (2023).
7. Kroeze, C., Dumont, E. & Seitzinger, S. P. New estimates of global emissions of N2O from rivers and estuaries. *Environ. Sci.* **2**, 159–165 (2005).
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. Anderson, E. P. *et al.* Understanding rivers and their social relations: A critical step to advance environmental water management. *WIREs Water* **6**, e1381 (2019).
10. 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).
11. 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).
12. 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).
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. 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).

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

17. Stadler, M. & del Giorgio, P. A. Terrestrial connectivity, upstream aquatic history and seasonality shape bacterial community assembly within a large boreal aquatic network. *ISME J.* **16**, 937–947 (2022).

18. 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).

19. 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).

20. Ruiz-González, C., Niño-García, J. P. & del Giorgio, P. A. Terrestrial origin of bacterial communities in complex boreal freshwater networks. *Ecol. Lett.* **18**, 1198–1206 (2015).

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

22. 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).

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

24. 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.

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

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

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

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. Parks, D. H. *et al.* Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nat. Microbiol.* **2**, 1533–1542 (2017).

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

31. 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).

32. 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).

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

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

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

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

37. 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.

38. 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).

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

40. 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).

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

42. 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).

43. 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).

44. 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).

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

46. Cory, R. M., Crump, B. C., Dobkowski, J. A. & Kling, G. W. Surface exposure to sunlight stimulates CO2 release from permafrost soil carbon in the Arctic. *Proc. Natl. Acad. Sci. U. S. A.* **110**, 3429–3434 (2013).

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

48. 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).

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

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

51. 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).

52. 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).

53. 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).

54. 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).

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

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

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

58. 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).

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

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

61. Jr, N. H. *Water and the West: The Colorado River Compact and the Politics of Water in the American West*. (2009).

62. Thompson, L. R. *et al.* A communal catalogue reveals Earth's multiscale microbial diversity. *Nature* **551**, 457–463 (2017).

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

64. *U. S. Geological Survey. Book 9, Chapter A4 Collection of water samples*. *in U.S. Chapter A4. Collection of water samples* vols 09-A4 231 http://pubs.er.usgs.gov/publication/twri09A4 (2006).

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

66. Lee, C. J. & Henderson, R. J. Tracking water quality in U. S. streams and rivers. https://nrtwq.usgs.gov/nwqn/#/ (2020).

67. Crump, B. C., Kling, G. W., Bahr, M. & Hobbie, J. E. Bacterioplankton Community Shifts in an Arctic Lake Correlate with Seasonal Changes in Organic Matter Source. *Appl. Environ. Microbiol.* **69**, 2253–2268 (2003).

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

69. 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).

70. 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).

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

72. Katie R. Willi & Matthew R. V. Ross. Genome Resolved Open Watersheds Database (GROWdb) Geospatial Data Puller GitHub. (2023).

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

74. 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).

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

76. 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.

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

78. 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).

79. 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).

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

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

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

83. 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).

84. 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).

85. 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).

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

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

Data accessibility

The data underlying GROWdb are accessible across multiple platforms to ensure many levels of data use and 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 in this manuscript 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 content 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)[2,23] data portal, the Department of Energy's Systems Biology Knowledgebase (KBase)[3], and a GROW specific user interface released here, GROWdb Explorer. Each platform provides different ways to interact with data in the 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, other data collected on the same samples, and standard analysis results, following Findable, Accessible, Interoperable, and Reusable (FAIR) data practices[30].
- *KBase* GROWdb is a publicly available collection within KBase[3], 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 analyses using KBase's 500+ analysis tools, in a point and click format.
- *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 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.

Code availability

All scripts involved with microbial data generation, processing, curation, and visualization are available on GitHub (https://github.com/jmikayla1991/Genome-Resolved-Open-Watersheds-database-GROWdb/tree/main). Code for geospatial analysis and GROWdb Explorer are available on GitHub (https://github.com/rossyndicate/GROWdb).

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). BCC, TB, and SPG were partially supported by U.S National Science Foundation awards DEB1840243, EAR1836768, and DEB1457794. 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 at University of Colorado 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).

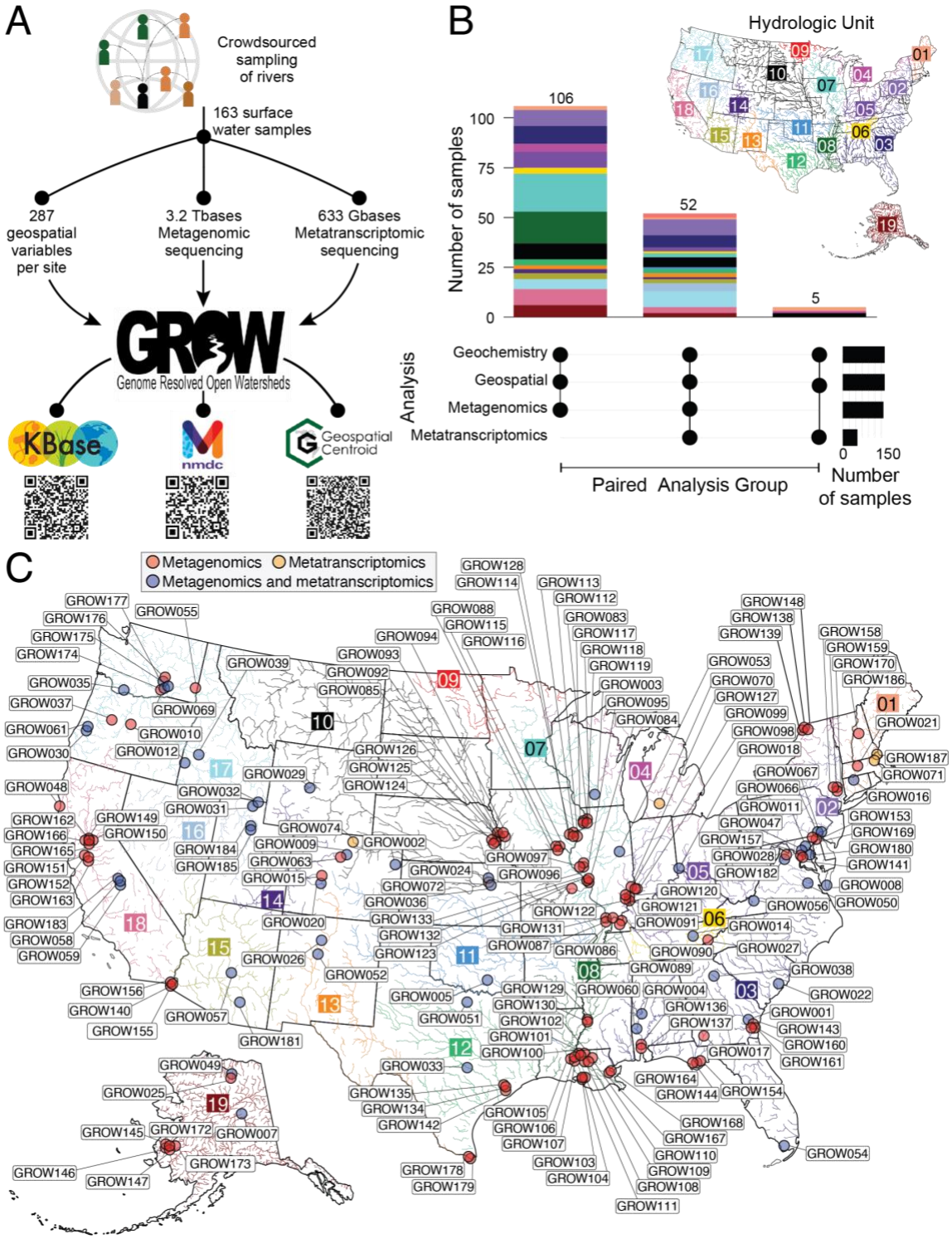


Figure 1. Distributed sampling and sequencing of rivers enabled the construction of the Genome Resolved Open Watersheds (GROW). A) Workflow denoting the number of samples and resulting datasets made up of geospatial and microbiome (metagenomics, metatranscriptomics) data. GROWdb

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).

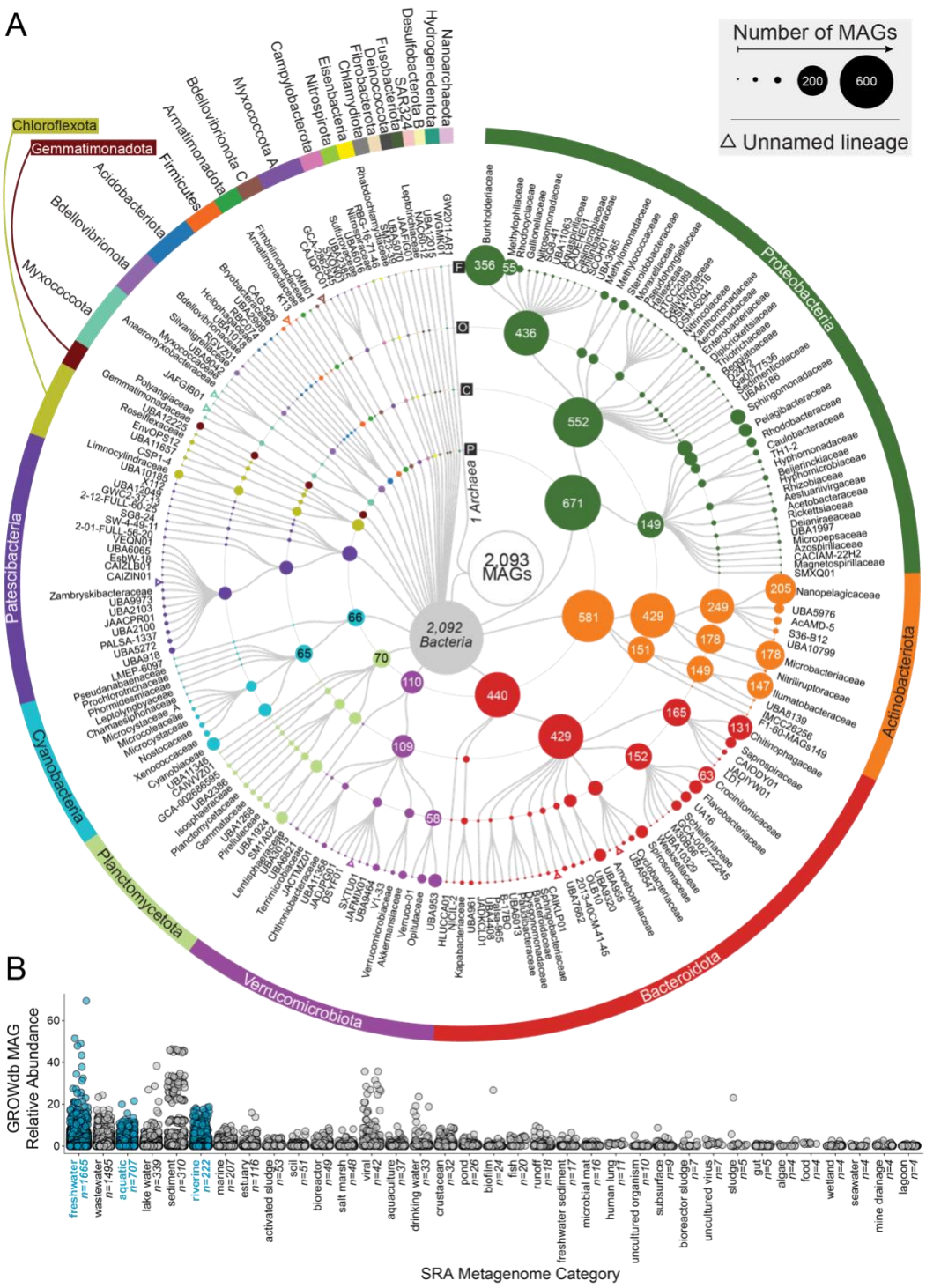


Figure 2: Taxonomic diversity of 2,093 unique surface water metagenome assembled genomes (MAGs) in GROWdb. A) Cladogram shows GROWdb MAGs taxonomy with each sequential ring noting taxonomy level (Phylum, P; Class, C; Order, O; Family, F). Circle size indicates the number of

genomes within a given taxonomy level and is further noted by MAG number inside the circle when sampling at that taxonomic position exceeds 50 MAGs sampled. Colors highlight phylum level taxonomy denoted on 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 266,764 metagenomes in the Sequence Read Archive 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.

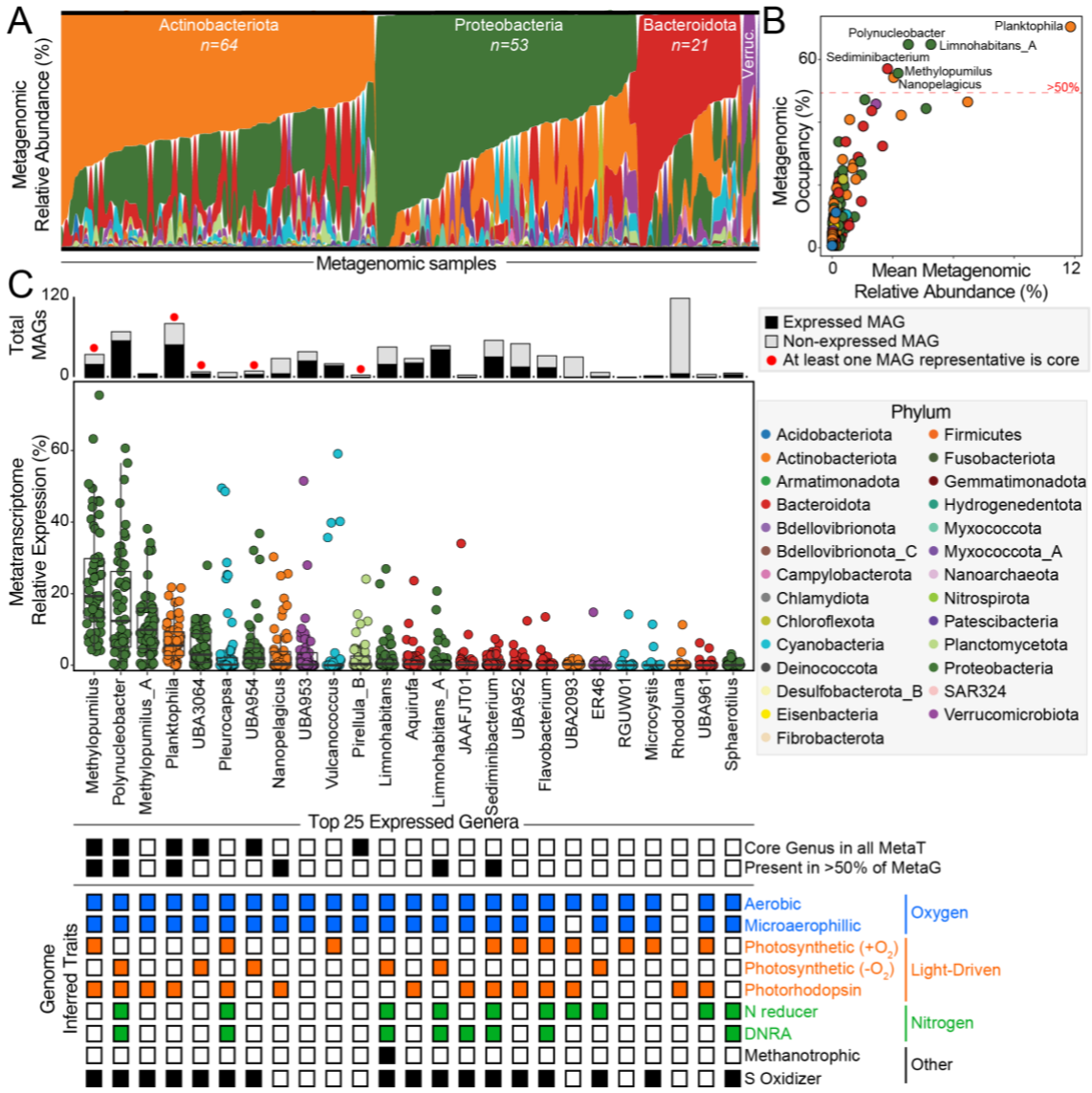


Figure 3: Core lineages and functions across river microbiomes. A) Phyla metagenomic relative abundance across samples, with each sample organized by the most dominant phyla from top to bottom along y-axis. Samples grouped by dominant phyla along the x-axis. Actinobacteriota, Proteobacteria, Bacteroidota, and Verrucomicrobiota (Verruc.) phyla are the most dominant 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 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 as defined as having gene expression in every sample. For each of the top 25 expressed genera, black boxes represent those that were detected in 100% of metatranscriptomes (core genera) and >50% of metagenomes. 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).

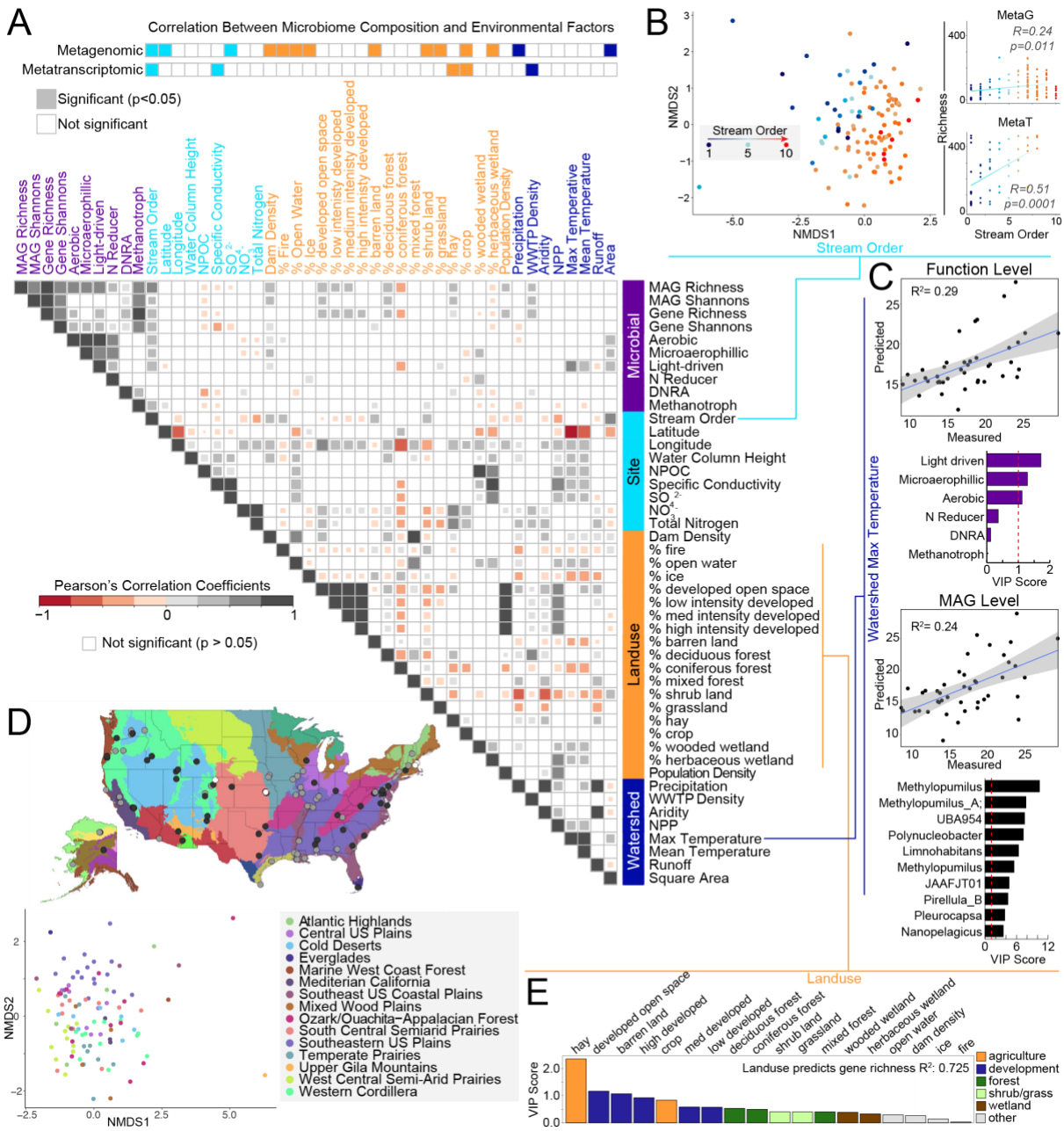


Figure 4: Patterns and drivers of river microbiome composition and function. A) Metagenomic and metatranscriptomic composition, function, and diversity were related to 36 selected site, land use, or watershed variables using Mantel tests (top, two rows). Below this was followed with pairwise comparisons using Pearson's correlation (heatmap in A). Variables are colored by category including microbial (purple), site or local (light blue), land use (orange) and watershed metrics (dark blue). For pairwise comparisons of microbial data, metatranscriptomic metrics were used for diversity and function

abundance calculations. B) Microbial community diversity was significantly associated with 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 predicted 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.

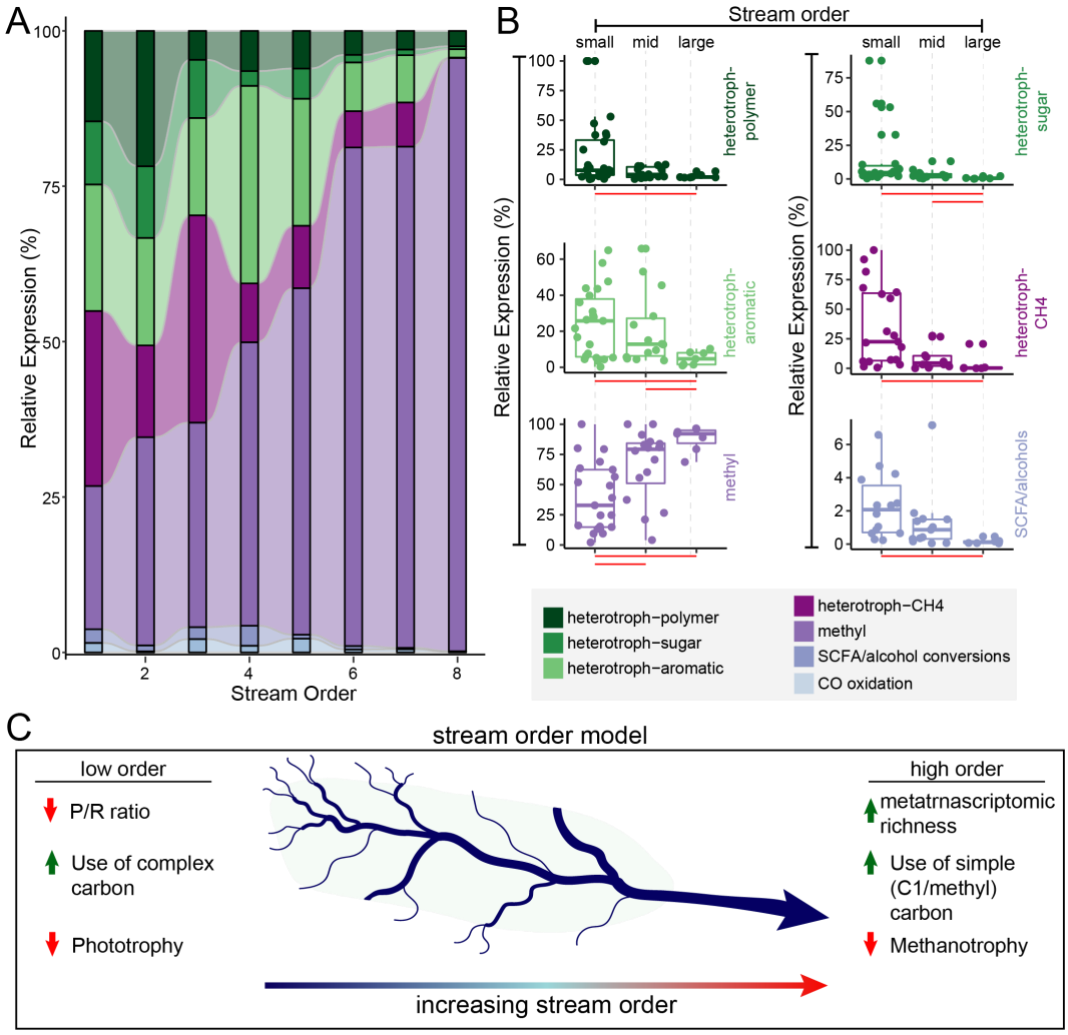
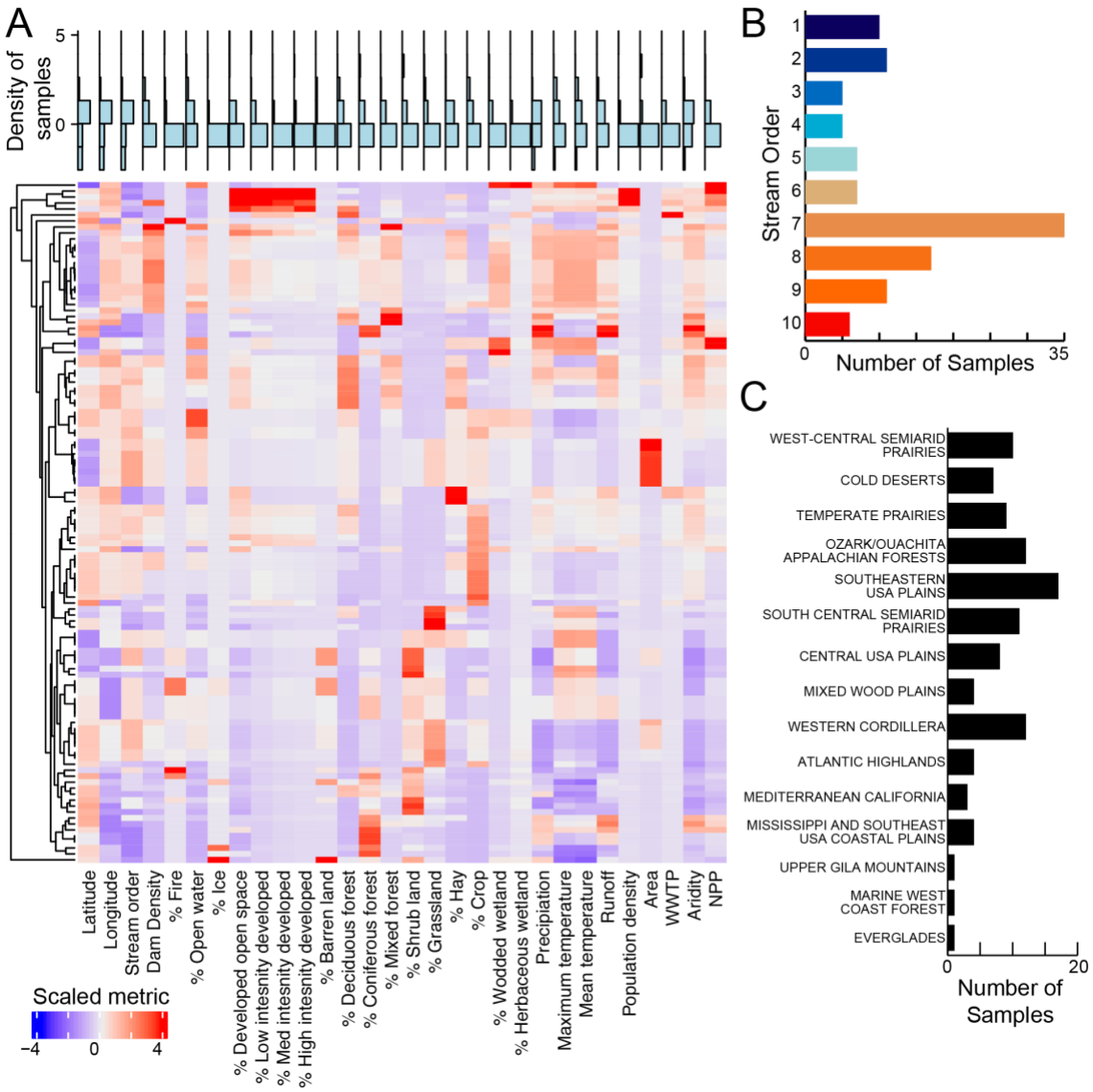
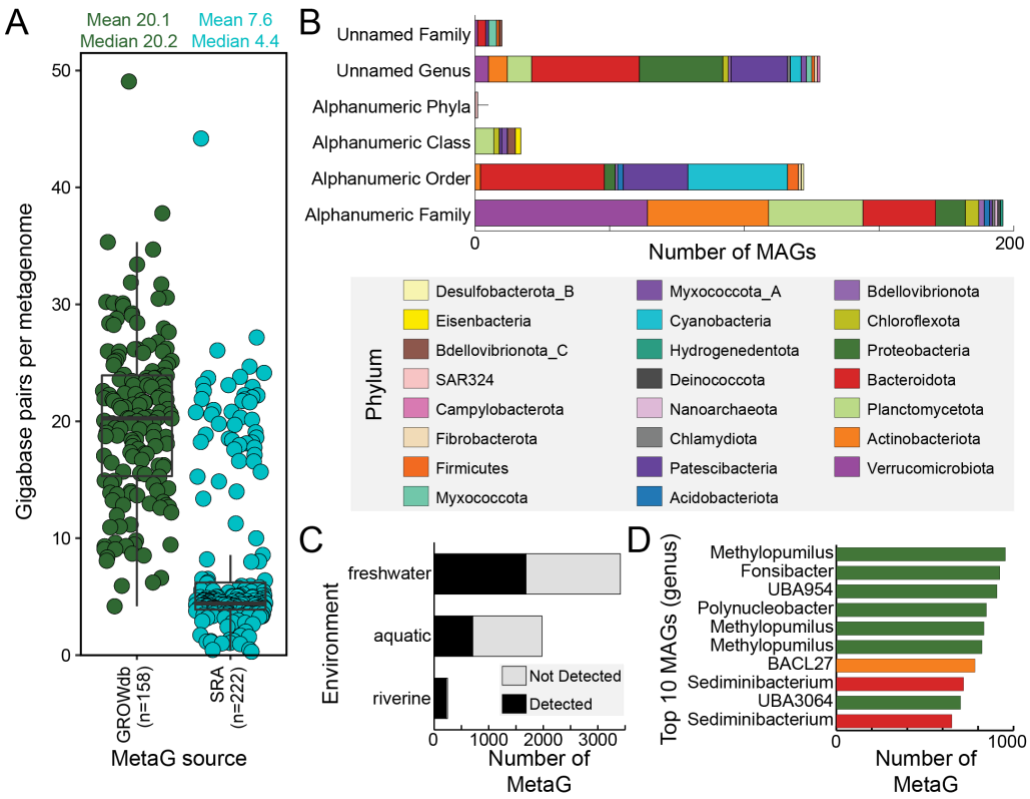


Figure 5: Microbial lifestyle and carbon utilization are structured along a stream order gradient. A) Alluvial plot shows the relative expression of microbial lifestyles (defined in methods) across stream order gradient. B) One dimensional scatterplots correspond to data in A, with each point representing 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.

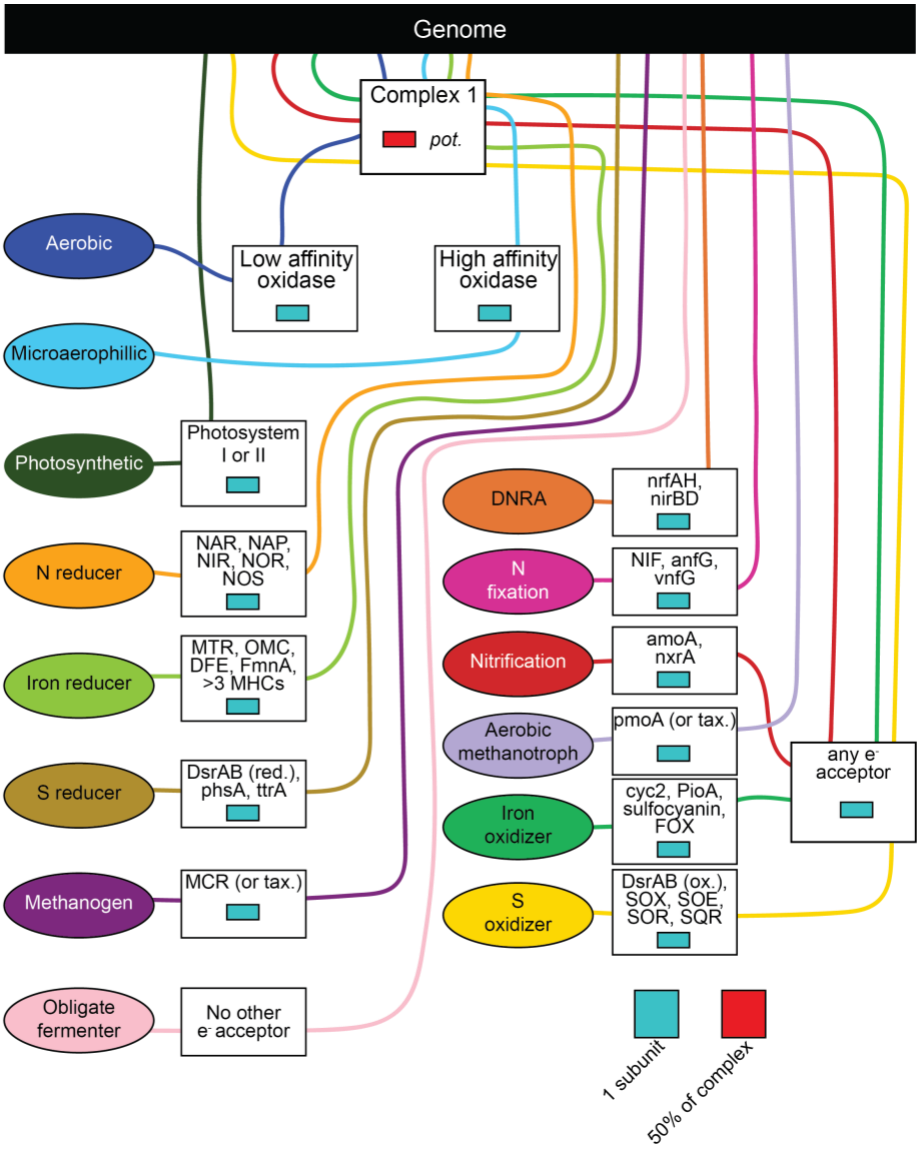
Extended Data Figures



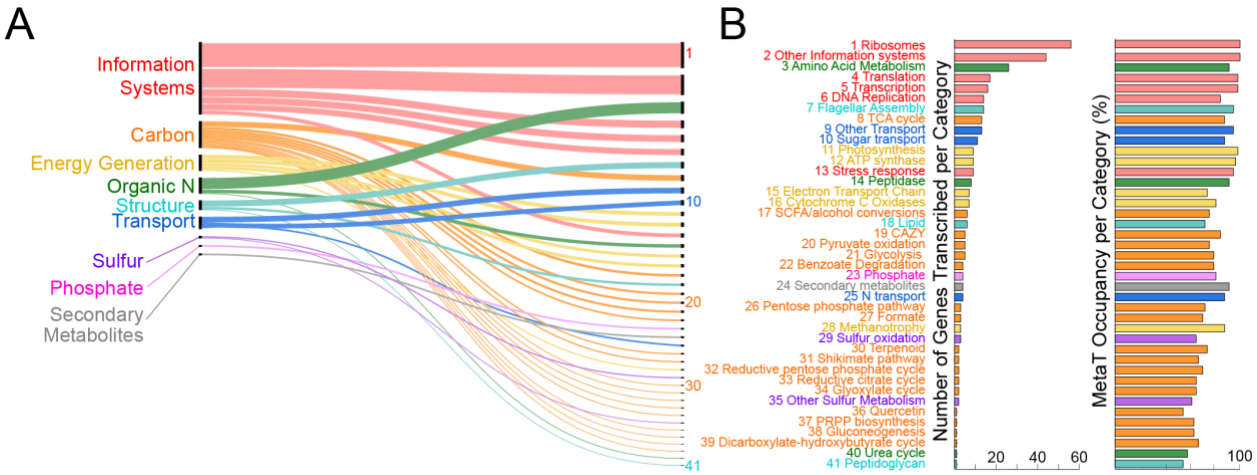
Extended Data Fig. 1: Distribution of river characteristics sampled in GROWdb. A) Heatmap of geospatial parameters sampled in GROWdb, where columns are environmental variables and rows are corresponding samples within GROWdb. Each variable has been scaled by subtracting the vector mean for each variable 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.



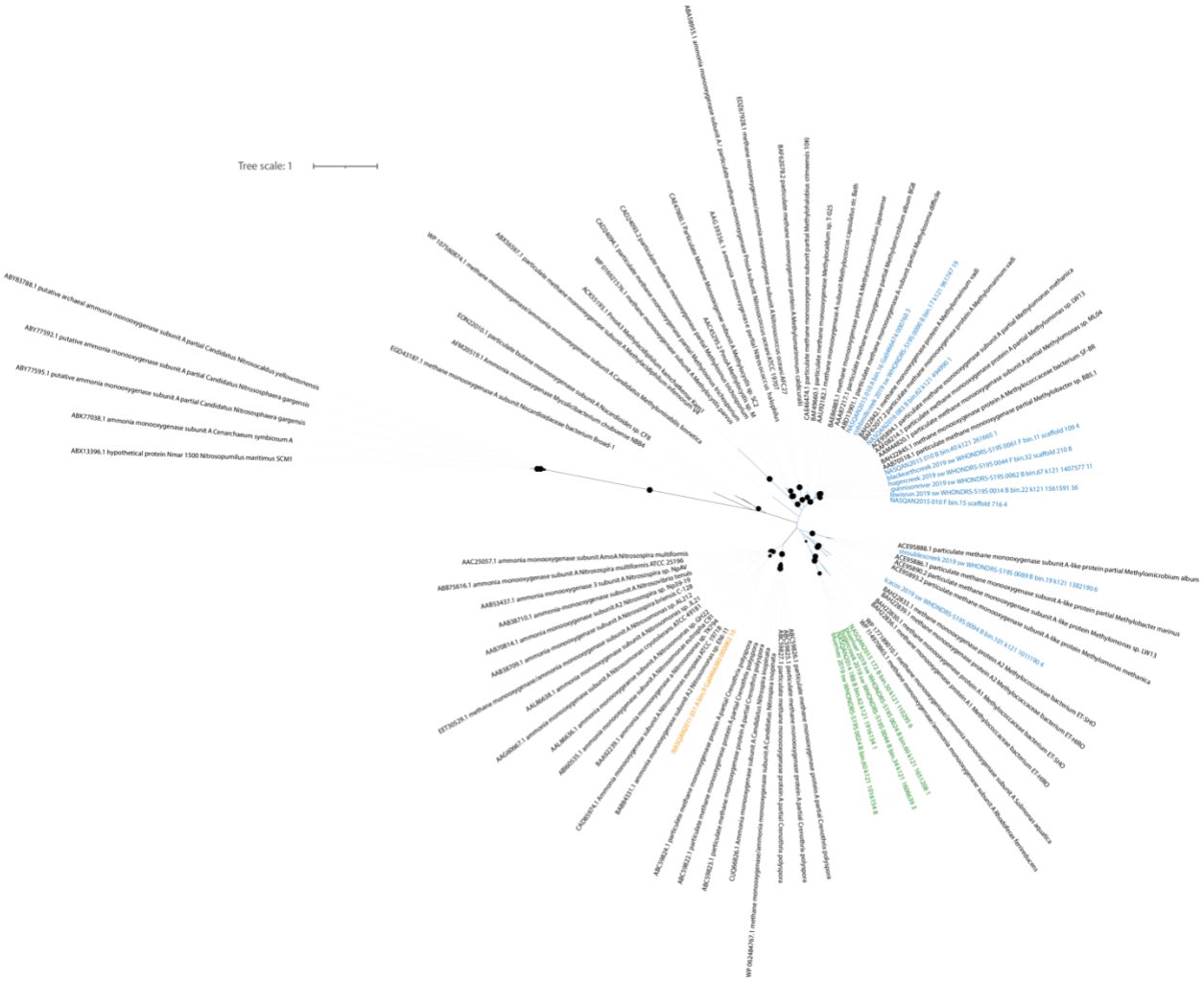
Extended Data Fig. 2: GROWdb comparison to other omics data sources. A) Sequencing depth comparison of GROWdb metagenomes (n=158) to SRA metagenomes classified as riverine (n=222) shows 3x increase in average sequencing depth for GROWdb. Each point represents a single metagenome, with mean and median values listed at the top of the graph. B) Stacked bar chart shows novelty of GROWdb MAGs when compared to GTDB (r207). Each MAG was placed at the highest level of novelty, with no assignment within a taxonomic level (e.g., unnamed family or genus) being highest level of novelty and alpha numeric identifiers being the second highest (e.g., UBA lineages). Bars are colored by Phylum. C) Stacked bar chart shows the proportion of SRA metagenomes that a GROWdb species was detected (black) or not detected (grey) within an SRA 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).

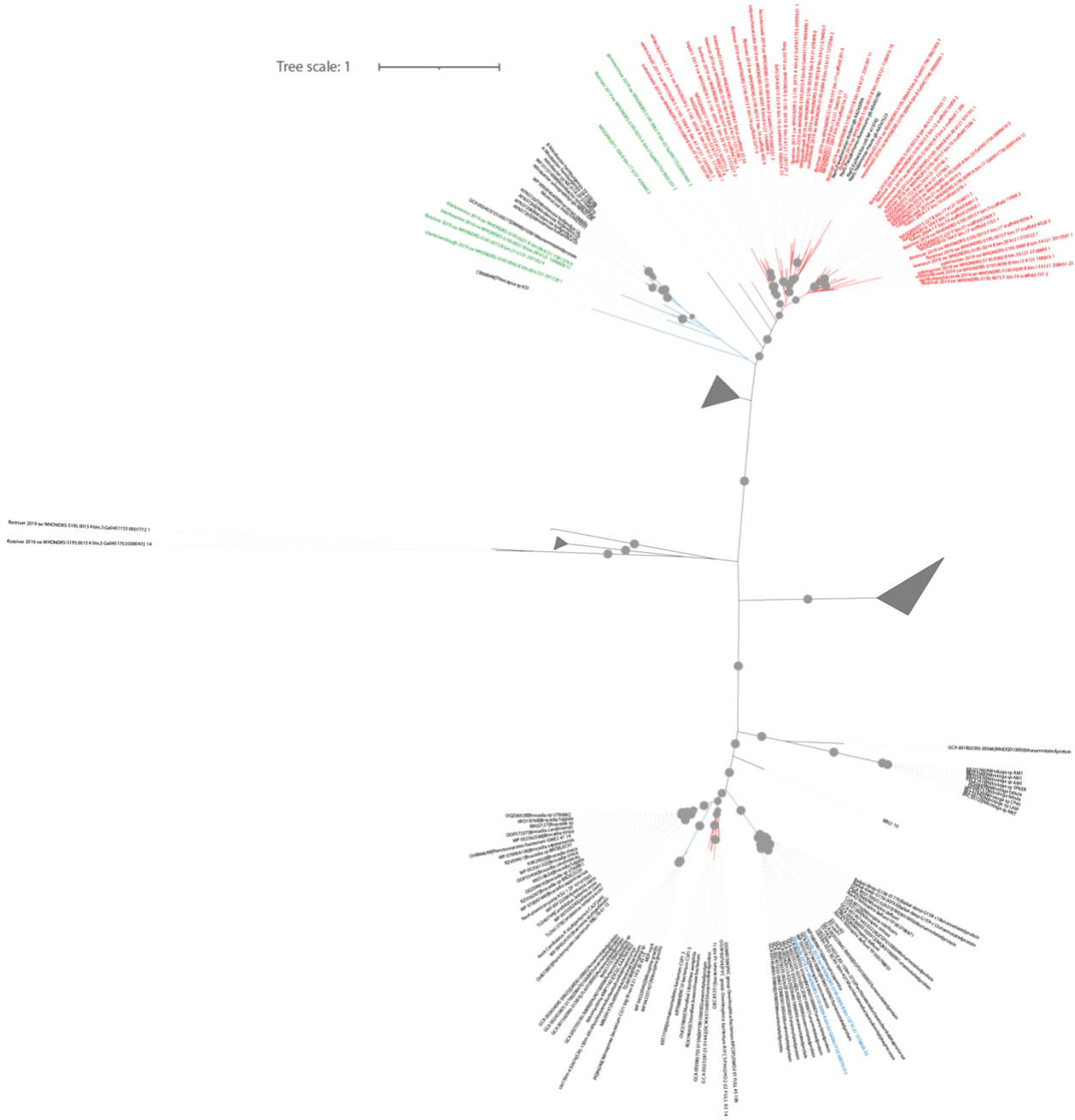


Extended Data Fig. 3: Metabolic trait assignment ruleset. Each trait is defined by a set of genes and the percent of genes required for that function. Lines flow from the genome (top black box) to traits (ovals), passing through boxes of gene requirements to be consider TRUE for that particular trait.

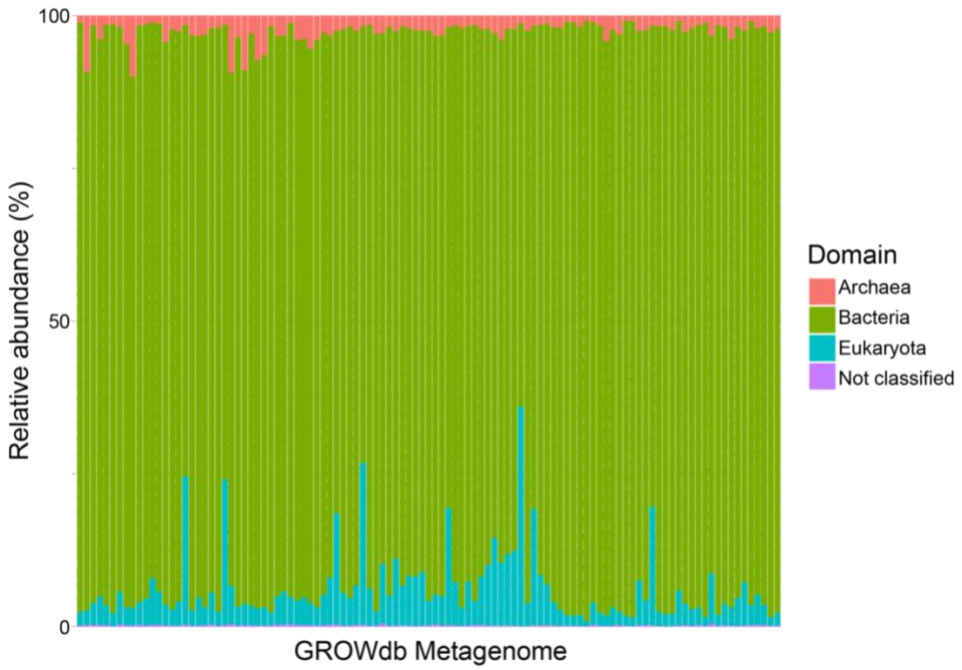


Extended Data Fig. 4: Gene level expression across rivers. Genes detected in more than 50% of metatranscriptomes, with gene functions (n=365) grouped by broad categories (n=9, A) and refined to subcategories (n=41, B). Thickness of lines and line order in A show the number of functions within a particular 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.

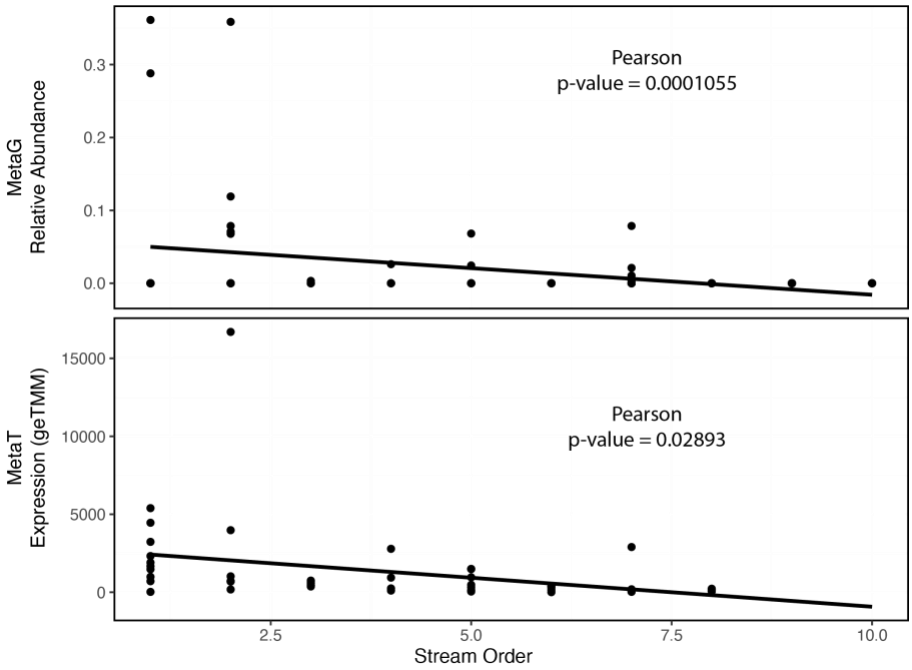

Tree scale: 1
ABY83788.1 putative archaeal ammonia monooxygenase subunit A partial Candidatus Nitrosocaldus yellowstonensis
ABY77592.1 putative ammonia monooxygenase subunit A partial Candidatus Nitrosophaera gargensis
ABY77595.1 putative ammonia monooxygenase subunit A partial Candidatus Nitrosophaera 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
AAB53437.3 ammonia monooxygenase 3 subunit A Nitrosospira sp. Nsp14
AAB38710.1 ammonia monooxygenase subunit A Nitrosomonas sp. Nsp10-15
AAB70814.1 ammonia monooxygenase subunit A Nitrosomonas sp. C-128
AAB38709.1 ammonia monooxygenase subunit A Nitrosomonas sp. Nsp5
AAL86508.1 ammonia monooxygenase subunit A Nitrosomonas cryotolerans ATCC 49181
AAL86509.1 ammonia monooxygenase subunit A Nitrosomonas sp. AL212
AAB60667.1 ammonia monooxygenase subunit A Nitrosomonas sp. Is79A3
AAB60035.1 ammonia monooxygenase subunit A Nitrosomonas europaea ATCC 19718
AAB32281.1 ammonia monooxygenase subunit A Nitrosomonas sp. Is79A3
AAB81031.1 ammonia monooxygenase subunit A Nitrosomonas nitrosa strain Nm90
NAGDAN2015 010 F bin.15 scaffold 716.4
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
WP 052484767.1 methane monooxygenase/ammonia monooxygenase subunit A Methylococcaceae bacterium (T-HIRO)
WP 052484769.1 methane monooxygenase/ammonia monooxygenase subunit A Methylococcaceae bacterium (T-SHO)
WP 052484771.1 methane monooxygenase/ammonia monooxygenase subunit A Methylococcaceae bacterium (T-HIRO)
WP 052484773.1 methane monooxygenase/ammonia monooxygenase subunit A Methylococcaceae bacterium (T-SHO)
WP 052484775.1 methane monooxygenase/ammonia monooxygenase subunit A Methylococcaceae bacterium (T-HIRO)
WP 052484777.1 methane monooxygenase/ammonia monooxygenase subunit A Methylococcaceae bacterium (T-HIRO)
WP 052484779.1 methane monooxygenase/ammonia monooxygenase subunit A Methylococcaceae bacterium (T-HIRO)
WP 052484781.1 methane monooxygenase/ammonia monooxygenase subunit A Methylococcaceae bacterium (T-HIRO)
WP



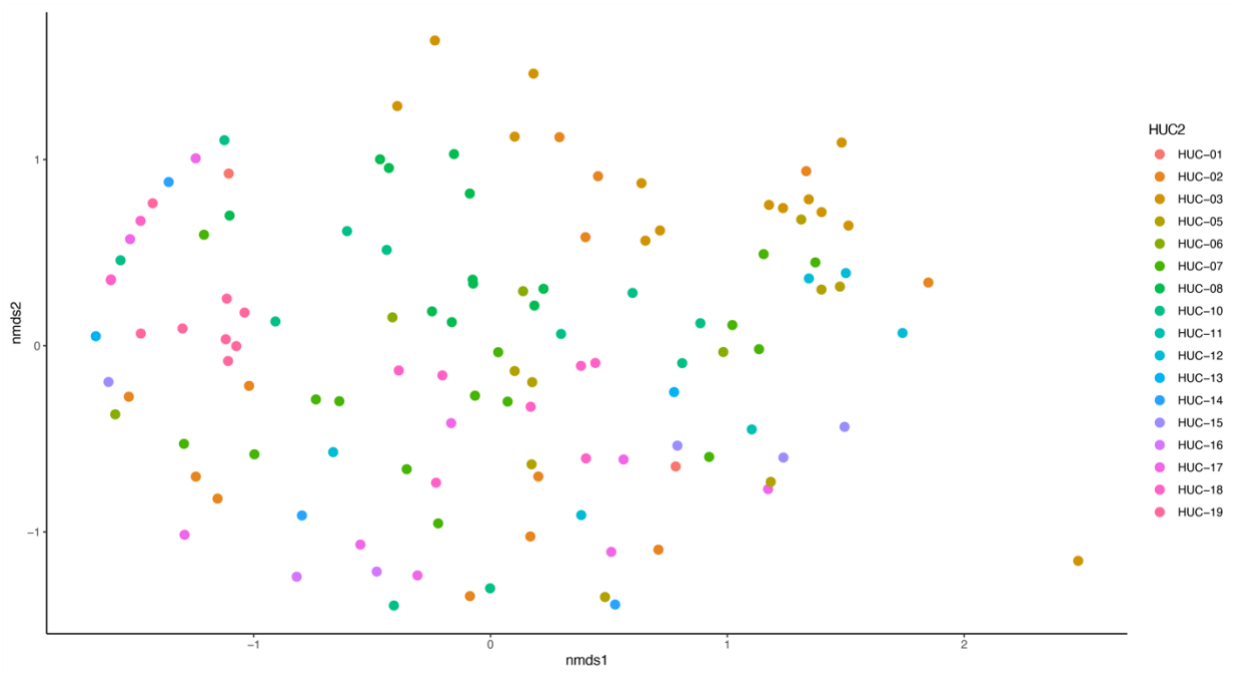
Extended Data Fig. 6: Phylogenetic analysis of nitrite oxidoreductase (nxr) and nitrate reductase (nar) is shown. GROWdb sequences colored by putative substrate, with nitrate (red), nitrite (blue), and unknown (green), while reference sequences are shown in black text. Bootstraps >80 are shown with closed grey circles at nodes. Tree files are available on Zenodo (https://doi.org/10.5281/zenodo.8173287).



Extended Data Fig. 7: Stacked bar chart of the singleM profiles of GROWdb metagenomic reads, with bars colored by domain. By domain, the most reads are assigned to the Bacteria (mean=91.1%), followed by Eukaryota (mean=6.1%), Archaea (mean=2.6%), and Unknown (mean=0.2%).



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



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

Extended Data Files

Extended data file 1: Excel file.

Tab 1) Table summary of river microbiome studies detailing title, authors, and microbiome methods used. Highlighted studies are genome-resolved multi-omics like GROWdb, with only 7 rivers surveyed.

Tab 2) Metagenome and metatranscriptome information.

Tab 3) Geospatial data (n=287 variables) for each sample.

Tab 4) List of datasets included in geospatial analysis.

Extended data file 2: Excel file.

Tab 1) Table summary of GROWdb MAGs including taxonomy and quality information.

Tab 2) Table summary of GROWdb novelty assigned by GTDB-Tk.

Tab 3) GROWdb MAG relative abundance across metagenomes.

Tab 4) Core analysis of GROWdb genera including metagenomic and metatranscriptomic occupancy and mean abundance/expression.

Extended data file 3: Excel file.

Tab 1) Annotation summary of GROWdb MAGs from DRAM.

Tab 2) GROWdb whole-genome trait calls.

Tab 3) Functional genes used to make whole-genome trait calls.

Tab 4) Output of Resistance Gene Identifier containing list of predicted ARGs

Tab 5) Metatranscriptomic expression of identified ARGs by category.