**Title:** Meta-metabolome ecology reveals that geochemistry and microbial functional potential are linked to organic matter development across seven rivers

**Authors:** Robert E. Danczak[a,*], Amy E. Goldman[a], Mikayla A. Borton[b], Rosalie K. Chu[c], Jason G. Toyoda[c], Vanessa A. Garayburu-Caruso[a], Emily B. Graham[c], Joseph W. Morad[d], Lupita Renteria[a], Jacqueline R. Hager[a], Shai Arnon[e], Scott Brooks[f], Edo Bar-Zeev[e], Michael Jones[f], Nikki Jones[f], Jorg Lewandowski[g,h], Christof Meile[i], Birgit M. Mueller[g], John Schalles[j], Hanna Schulz[g,h], Adam Ward[k], James C. Stegen[a]

[a]Pacific Northwest National Laboratory, Washington, USA

[b]Colorado State University, Colorado, USA

[c]Environmental Molecular Sciences Laboratory, Washington, USA

[d]New Jersey Department of Environmental Protection, New Jersey, USA

[e]Ben-Gurion University of the Negev, Israel

[f]Oak Ridge National Laboratory, Tennessee, USA

[g]Leibniz Institute of Freshwater Ecology and Inland Fisheries, Germany

[h]Humboldt University, Berlin, Germany

[i]University of Georgia, Georgia, USA

[j]Creighton University, Nebraska, USA

[k]Oregon State University, Oregon, USA

*Corresponding Author
Email: robert.danczak@pnnl.gov
Address: 902 Battelle Blvd.
P.O. Box 999, MSIN J4-18
Richland, WA 99352

**Abstract**

Rivers receive substantial dissolved organic matter (DOM) input from the land and transport it to the ocean. As DOM travels through watersheds, it undergoes biotic and abiotic transformations that impact biogeochemical cycles and any subsequent $CO_2$ release into the atmosphere. While recent research has increased our mechanistic knowledge of DOM composition within watersheds, DOM development across broad spatial distances and within divergent biomes is under investigated. Here, we combined DOM characterization, geochemical analyses, and shotgun metagenomics to analyze samples from seven rivers ranging from the U.S. Pacific Northwest to Berlin, Germany. Initial analyses revealed that many DOM properties were distinguished by river type (e.g., wastewater, headwater) and that geochemistry often explained variation across rivers. At a global scale, analyses rooted in meta-metabolome ecology indicated that DOM was structured overwhelmingly by deterministic selection. When controlling for scale, however, analyses indicated that ecological assembly dynamics were again partially structured by river type. Finally, microbial analyses revealed that many riverine microbes from our systems shared core metabolic functional potential while differing in peripheral capabilities in across the rivers. Further analysis of the carbon degradation potential for recovered metagenomically assembled genomes indicated that the sampled rivers had strong taxonomically conserved niche differentiation and that carbon degradation potential diversity was significantly related to organic matter diversity. Together, these results help us uncover interconnections between the development of DOM, riverine geochemistry, and microbial functional potential.

## 1. Introduction

Evaluating the principles underlying the composition of river corridor dissolved organic matter (DOM) has been an ongoing, cross-disciplinary endeavor (e.g., ecology, organic matter characterization, biogeochemistry, climatology) (Danczak et al., 2021; Gómez-Gener et al., 2021; Hu et al., 2022; Kaiser et al., 2017; Liu et al., 2019; Xenopoulos et al., 2021). This is because a significant quantity of organic matter is transported through rivers and undergoes combinations of abiotic and biotic transformations that eventually result in the release of $CO_2$ into the atmosphere (Cole et al., 2007; Drake et al., 2018; Gómez-Gener et al., 2021; Graham et al., 2018; Regnier et al., 2013; Stegen et al., 2018). By identifying the principles responsible for promoting and constraining DOM composition and incorporating them into various modeling capabilities, we can improve our ability to evaluate watershed function under future conditions (Song et al., 2020; Wohl et al., 2017; Zarnetske et al., 2018).

Microbial communities are one of the more commonly studied factors responsible for influencing DOM dynamics because they can directly interact with various DOM components. Most studies looking at the interconnections between microbial communities and riverine DOM examine a single watershed or reach (Garayburu-Caruso et al., 2020; Graham et al., 2018; Grunert et al., 2021; Osterholz et al., 2016; Saarela et al., 2022; Stegen et al., 2016a, 2018). Within the Columbia River Basin, researchers revealed that mixing dynamics and thermodynamic controls were the key intersections between DOM and microbiology (Graham et al., 2018; Stegen et al., 2016a, 2018). Salinity was an overarching control on both the DOM composition and microbial community structure within the Delaware Estuary, suggesting that external environmental pressures can outweigh the impacts of either assemblage (Osterholz et al., 2016). Studies investigating

respiration rates have revealed that DOM thermodynamics play a role in impacting microbial activity, but only when carbon is a limiting factor (Garayburu-Caruso et al., 2020).

Given the heterogeneity of the river corridors above, holistic studies that incorporate information from many different sampling sites are necessary to uncover these principles. Looking at 37 different aquatic environments, Kellerman et al. (2018) revealed that the age of organic matter was consistently related to DOM composition indicating the importance of degradation state. In support of these findings, D'Andrilli et al., (2015) demonstrated that organic matter lability was an important differentiating factor across various aquatic environments, revealing that glacial DOM was the most labile, followed by marine DOM, and finally freshwater. Hu et al. (2022) studied aquatic microcosms across temperature and nutrient gradients and observed that differential assembly processes affected the active and inactive components of DOM, with significant implications for inherent functional traits of DOM molecules. By further expanding both the range of samples and by integrating multiple data streams, we can better uncover important factors impacting DOM composition.

While these studies reveal different factors influencing DOM composition and microbial community structure, these components are often analyzed separately. The recently proposed analytical framework known as 'meta-metabolome ecology' represents a potential avenue to develop connections between these data types (Danczak et al., 2020). This analysis treats DOM as an assemblage of molecular formulas subject to processes akin to ecological community assembly processes. By analyzing DOM data this way, researchers can identify transferrable biogeochemical principles and connect biological and chemical data. For example, the use ecological null modeling

on DOM helped develop the concept of thermodynamic redundancy, a parallel of functional redundancy observed in microbial communities but for DOM (Danczak et al., 2021). The application of these ecological tools has begun to reveal connections between DOM and microbial communities as well. Similar analyses of both a riverine DOM assemblage and microbial community revealed a lack of selective interconnections, contrary to expectation (Danczak et al., 2020). A combination of DOM null modeling and functional trait analysis indicated that deterministic assembly is predicted to increase under situations of increased global change (e.g., changing temperature/nutrient conditions) (Hu et al., 2022). Turning away from null modeling, functional diversity of DOM assemblages and microbial communities within a freshwater lake revealed significant covariation indicating a potentially underappreciated role in the global carbon cycle (Tanentzap et al., 2019). A study analyzing DOM at a global scale may help reveal transferable principles concerning the behavior of organic matter and is needed because many studies are focused on specific freshwater ecosystems with inconsistent patterns.

Here, we investigated samples from seven rivers spanning a broad geographic extent from the Middle East to the US Pacific Northwest, as part of the Worldwide Hydrobiogeochemical Observation Network for Dynamic Rivers (WHONDRS) consortium. By combining DOM characterization, geochemical analysis, and shotgun metagenomic sequencing partnered to ecological analyses, we assessed environmental interactions which may explain global, river corridor DOM variation and integrate prior observations. In doing so, we sought to characterize the microbes that interact with and alter river corridor DOM. The primary goal of these analyses is to obtain a cross-ecosystem and transferrable understanding of river DOM development and to observe connections between DOM and the microbes that consume it. Ultimately, we observed

that the underlying geochemical and lithological conditions appear to be more strongly connected to DOM composition than other factors and that globally distributed microbial species may have conserved niches regarding carbon degradation.

## 2. Materials and Methods

*2.1 Site Descriptions.* This study obtained samples from the seven rivers spanning different biomes and flow regimes (Table 1, Figure 1). These rivers were the Altamaha River (Georgia, USA), Columbia River (Washington, USA), East Fork Poplar Creek (Tennessee, USA), Erpe River (Germany), HJ Andrews – Watershed 1 (Oregon, USA), Jordan River (Israel), and Nisqually River (Washington, USA). The Altamaha River drains a large portion of the state of Georgia to the Atlantic Ocean and has a watershed that extends over approximately 36,000 km2, encompassing the Northern Piedmont and the Coastal Plain (Takagi et al., 2017). The site experiences semidiurnal tidal variations in water levels, with a tidal amplitude on the order of 1-2 m, and occasionally experiences periods of salinity during low flow periods of October. The Columbia River is a gravel bedded river with some riparian vegetation and experiences variable flow associated with dam activity (Arntzen et al., 2006; Stegen et al., 2016a; Villa et al., 2020). The East Fork Poplar Creek has a streambed of gravel, cobble, and bedrock and is surrounded primarily by forest (Demers et al., 2018). This site was historically impacted by mercury contamination because of nuclear weapon development (Brooks and Southworth, 2011; Demers et al., 2018; Loar et al., 2011). The Erpe River is an urban lowland stream that consists of up to 80% treated wastewater originating from an upstream wastewater treatment plant (Mueller et al., 2021). The watershed at HJ Andrews is a shallow, forested, headwater stream that experiences diel fluctuations driven by evapotranspiration (Danczak et al., 2021). The Jordan River is a karstic fed

stream found within the northern section of the Syrian African Rift Valley and is subject to discharge fluctuations due to groundwater pumping (Goldman et al., 2019). Finally, the Nisqually River is low-order stream fed by that originates in Mount Rainier and experiences diel fluctuations in discharge due to glacial freeze/thaw cycles (Curran et al., 2016). Additional information and site photos for each river can be found in associated data packages published on ESS-DIVE (see Section 2.8 for complete citations).

*2.2 Sample Collection and Coordination.* Samples from the seven rivers were collected between July and October 2018 as part of a WHONDRS initiative to study diel river dynamics (Stegen and Goldman, 2018). As such, the sampling scheme was similar to the methodology described in Danczak et al., 2021 and was originally devised to investigate temporal patterns despite this study focusing on spatial variation. Three positions separated by ~4m were selected at each river corridor to collect pore water samples. Approximately 20 mL was collected from each of these locations every 3 h over a 48 h period. Surface water was additionally collected in triplicate from the same position as the middle pore water sampling site. In total, 690 total samples were collected over roughly 17 time points. Surface water was collected using a 60 mL syringe through Teflon tubing while the pore water was collected using a syringe attached via Teflon tubing to a 30 cm long stainless-steel sampling tube (MHE Products, MI, USA) with a slotted screen across the bottom ~5 cm. One sampling tube was installed to 30 cm depth at each pore water sampling location; these tubes sat in place for around 15 min before the first sample was collected and they remained in place during the 48-h time course of sampling. Prior to sampling a given location, the syringe was flushed 3 times with the source water to ensure only the desired water was collected. All samples were filtered through a 0.2 μm Sterivex filter (Millipore, MA, USA). At each time point, one filter

was used for all three pore water samples, and a different filter was used for all three surface water samples. Water passing through the filters was collected for analysis using a needle attached to the filter and injected through a septum. During sampling, water temperature, approximate water stage, and pH were measured. Water samples for DOM analysis were injected into amber borosilicate glass vials. Samples for cations and anions were injected into clear borosilicate glass vials. Once collected, samples were stored in a cooler on blue ice until they could be frozen until they were processed in the lab.

*2.3 Geochemical Analyses.* As with the sampling scheme, the geochemical analyses are similar to the methods described in Danczak et al., 2021. In brief, anion concentrations were measured using a Dionex ICS-2000 ion chromatograph with AS40 autosampler using an isocratic method (guard column: IonPac AG18 guard, 4x50mm; analytical column: IonPac AS18, 4x250mm; suppressor: RFIC ASRS, 300 rmm, self-regenerating; suppressor current: 57 mA). Standards were made from Spex CertiPrep (Metuchen, NJ, 08840) 1000 mg/L anion standards and were diluted as follows: $NO_2$ in the range of 0.04 to 20 ppm, F in the range of 0.2 to 10 ppm, Cl and $SO_4$ in the range of 0.16 to 80 ppm, and $NO_3$ in the range of 0.12 to 60 ppm. Cation samples were prepared with nitric acid and were measured with a Perkin Elmer Optima 2100 DV ICP-OES with an AS93 auto sampler. Calibration standards were made with Ultra Scientific ICP standards (Kingstown, RI): P, Mg, Ca, K, and Na were diluted in the range of 5–4000 ppm, and Fe was diluted in the range of 0.5–400 ppm. Non-purgeable organic carbon (NPOC) was determined by a Shimadzu combustion carbon analyzer TOC-L CSH/CSN E100V with ASI-L auto sampler calibrated to a range of 0.5 to 10 ppm NPOC as C via potassium hydrogen phthalate (Nacalia Tesque, lot M7M4380). An aliquot of sample was acidified with 15% by volume of 2 N ultra-pure HCL and then sparged with carrier

gas for 5 min to remove inorganic carbon. The best 3 out of 4 injections replicates were averaged to get the final result.

*2.4 Dissolved organic matter sample analysis and data processing.* All samples were analyzed using a Fourier-Transform Ion Cyclotron Resonance mass spectrometer (FTICR-MS) following methods established previously (Danczak et al., 2021, 2020; Garayburu-Caruso et al., 2020). In brief, samples were standardized to a given carbon concentrations (NPOC 0.69 – 1.5 mg C/L), acidified to pH 2 with 85% phosphoric acid, and extracted with PPL cartridges. High-resolution mass spectra were collected from these samples using a 12 Tesla (12T) Bruker SolariX FTICR mass spectrometer (Bruker, SolariX, Billerica, MA) equipped with an electrospray ionization (ESI) source set to negative mode. BrukerDaltonik Data Analysis (version 5.0) was used to generate a list of m/z values from the raw spectra using the FTMS peak picker module. Chemical formulae were assigned using Formularity (Tolić et al., 2017) according to the Compound Identification Algorithm (Kujawinski and Behn, 2006; Minor et al., 2012; Tfaily et al., 2017) with the following settings: a signal-to-noise ratio >7, mass measurement error <0.5 ppm, and allowing for CHONSP (Formularity formula settings: O > 0 AND (N + S + P) < 6 AND S < 3 AND P < 2).

The R package ftmsRanalysis (Bramer et al., 2020; Bramer and White, 2019) was used to remove peaks that were outside the desired m/z range (200 m/z – 900 m/z) or had an isotopic signature, calculate a number of derived metrics (double-bond equivalent (DBE), modified aromaticity index (AIMod), nominal oxidation state of carbon (NOSC)), and organize the data (Hughey et al., 2001; Koch and Dittmar, 2006; LaRowe and Van Cappellen, 2011; Tfaily et al., 2015). Given the limitations associated with peak intensities, all intensities were converted to presence/absence

214 (Tfaily et al., 2017). Once converted to presence/absence, replicates for each sample were merged
215 such that a peak needed to be present in at least two of the three replicates for it to be considered
216 "present" in that sample. For each sample, we also calculated molecular formula richness as the
217 number of molecular formulas assigned in the respective sample.
218
219 *2.5 Meta-Metabolome Ecology.* Following the protocols established in Danczak et al., we
220 generated a transformation-weighted molecular characteristics dendrogram (TWCD) (Danczak et
221 al., 2020). The molecular characteristics dendrogram component of the dendrogram was generated
222 by first measuring the Euclidean distance between molecular formulas based on their derived
223 metrics (e.g., elemental composition, double-bond equivalents, modified aromaticity index, and
224 Kendrick's defect). These characteristic distances were then weighted by transformation distances
225 associated with the mass of each molecular formula. While described in detail elsewhere (Buser-
226 Young et al., 2023; Danczak et al., 2021, 2020; Fudyma et al., 2021; Graham et al., 2018), the
227 transformation analysis estimates the presence of putative biochemical transformations by
228 measuring the mass differences between identified peaks which are then related to a database of
229 known and commonly observed biochemical transformations. Relationships between peaks are
230 then established by measuring the number of transformations required for one peak to potentially
231 become another. These distances are normalized to a range between 0 and 1 and are used to down
232 weight the molecular characteristic distance. This transformation-weighted molecular
233 characteristic distance matrix is then used as input for an unweighted pair group method with
234 arithmetic mean (UPGMA) hierarchical cluster analysis. The output of this analysis represents
235 total potential relations that could exist across all DOM identified in this study, therefore, this
236 TWCD represents an approximation of global DOM relationships.

The TWCD was used primarily to perform a null modeling analysis known as β-nearest taxon index (βNTI) to estimate the ecological assembly processes impacting the DOM (Danczak et al., 2020; Stegen et al., 2015, 2013, 2012). To perform this analysis, a null distribution of 999 β-mean nearest taxon distance (βMNTD) values were generated and compared to the observed βMNTD value for a given pair of assemblages within each river (e.g., cross-river comparisons were not analyzed). Pairwise comparisons with $|\beta NTI| > 2$ indicate that deterministic processes were responsible for observed differences in metabolite composition. In contrast, pairwise comparisons with $|\beta NTI| < 2$ indicate that stochastic processes were responsible for observed differences in metabolite composition. Furthermore, the deterministic processes can be separated into two classes. When $\beta NTI > 2$, differences in metabolite composition are greater than would be expected by random chance (i.e., greater than the stochastic expectation). This is analogous to 'variable selection,' which occurs when deterministic processes drive divergence in composition between a pair of assemblages (Dini-Andreote et al., 2015; Stegen et al., 2015). When $\beta NTI < -2$, differences in metabolite composition are less than the stochastic expectation. This is analogous to 'homogeneous selection,' which occurs when deterministic processes drive convergence in composition between a pair of assemblages. We would like to highlight that while null modeling-was performed on each river independently, the same dendrogram was used in each analysis and was not pruned. This approach allowed us to take a more global perspective in assembly processes and potentially overcome some of the scale limitations observed in other meta-metabolome ecology studies (Danczak et al., 2023, 2021).

259 Using a similar approach to the microbial data above, we analyzed alpha diversity and beta
260 diversity using the diversity and vegdist/metaMDS functions, respectively, from the vegan R
261 package (v2.6.4). Additionally, we performed a distance-based redundancy analysis (dbRDA)
262 leveraging Jaccard dissimilarity to evaluate relationships between the meta-metabolome and
263 geochemical variables. First, we limited the effect of co-correlated variables by identifying them
264 using the vif.cca function and manually removing variables with high degrees of co-correlation
265 from our model; variables removed from the model were K, Na, and $SO_4$ Next, we performed
266 forward and reverse variable selection to identify the minimum set of geochemical variables that
267 could explain metabolome variation using the ordiR2step function. The final set of selected
268 variables were NPOC, Cl, Mg, TN, $NO_3$, F, P, Fe, and pH. Finally, we evaluated the significance
269 of the final dbRDA utilizing a permuted ANOVA and visualized the RDA with loadings.
270
271 *2.6 DNA Extraction and Sequencing.* DNA was extracted from filters using Zymobiomic's
272 Quick-DNA Fecal/Soil Microbe Microprep kit (cat# 6012) using the manufacturer's protocol and
273 the soil option. Metagenomic libraries were prepared using the Tecan Ovation Ultralow System
274 V2 and were sequenced on the NovaSEQ6000 platform on a S4 flow cell at Genomics Shared
275 Resource, Colorado Cancer Center, Denver, CO, USA.
276
277 *2.7 Metagenome Assembly, Binning, and Annotation.* All raw reads were trimmed for length and
278 quality using Sickle v1.33 (Joshi and Fass, 2011). To maximize bin recovery and quality, we used
279 multiple assembly methods which are denoted in each assembly and bin name as follows: (B)
280 megahit v1.1.1, (C) IDBA-UD 1.1.0, (D) metaspades v3.13.0, (E) bbcms followed by megahit
281 v1.1.1, and (F) subsample reads (25%) using bbtools followed by IDBA-UD 1.1.0 (Bushnell, 2020;

282 Li et al., 2015; Nurk et al., 2017; Peng et al., 2012). All assemblies were then individually binned
283 using Metabat2 v2.12.1 with default parameters to obtain microbial metagenome assembled
284 genomes (MAGs) (Kang et al., 2015).
285
286 MAG completion was estimated using CheckM v1.1.2 (Parks et al., 2015). To ensure only quality
287 MAGs were utilized for metabolic analyses, we discarded all MAGs that had completion <50%
288 and contamination >10%. This equates to "high quality" (HQ) bins and "medium quality" (MQ)
289 bin cutoffs outlined by the genome consortium standard. The resulting MAGs were manually
290 curated and then dereplicated using dRep with default parameters to result in a final set of 302
291 MAGs (>99% ANI which represents strain-level MAG distinctions) (Olm et al., 2017)
292 (Supplemental Table 1). To quantify MAG relative abundance in each temporal sample and
293 condition, trimmed metagenomic reads were mapped to the dereplicated MAG set using bbmap
294 (v38.70) at minid=95, and output as sam files which were converted to sorted bam files using
295 samtools (v1.9) (Bushnell, 2020; Danecek et al., 2021). We required reads to map to at least 75%
296 of a MAG in each sample, and second the MAG had to have at least 3x coverage in that sample.
297 To determine MAGs that had reads mapped to at least 75% of the MAG, we used CoverM (v0.3.2)
298 in genome mode to output MAGs that passed this threshold (–min-covered-fraction 75); MAGs
299 were considered present within a sample if they had mapping above this threshold (Woodcroft,
300 n.d.). To obtain reads per base for each MAG, we used CoverM (v0.3.2) in genome mode to output
301 reads_per_base (reads mapped/genome length), and from this calculated MAG coverage as
302 reads_per_base x 151 bp (Woodcroft, n.d.). MAGs were taxonomically classified and using the
303 Genome Taxonomy Database (GTDB) Toolkit v1.3.0 in August 2020 (Chaumeil et al., 2020).
304 MAG scaffolds were annotated using the DRAM v1.0 tool which uses PFAM (v33.1), KEGG

305 (v89.1), dbCAN (v9), MEROPS (v120), and VOGDB databases (https://vogdb.org/) for
306 annotations (Finn et al., 2016; Kanehisa et al., 2016; Rawlings et al., 2018; Yin et al., 2012). A
307 phylogenetic tree visualization was generated by 1) identifying ribosomal protein S3 (rps3) amino
308 acid sequences in each MAG by leveraging a Hidden Markov Model (HMM) from PFAM, 2)
309 aligning rps3 sequences using MUSCLE (Edgar, 2004), 3) trimming the MUSCLE alignment
310 using trimAl (Capella-Gutiérrez et al., 2009), 4) using the alignment to generate a phylogenetic
311 tree using FastTree (Price et al., 2010), and 5) finally visualizing the tree using the ggtree R
312 package (Yu et al., 2017). Additional analyses were performed on all MAGs using the dbCAN
313 standalone pipeline.
314
315 Using the MAG coverage data and the CAZyme count information obtained via the dbCAN
316 analyses, we performed alpha diversity analyses in each sample (diversity, vegan package v2.6.4;
317 Oksanen et al., 2019). Beta-diversity was measured using Bray-Curtis dissimilarity (vegdist, vegan
318 package v2.6.4) and was visualized using non-metric multidimensional scaling (NMDS;
319 metaMDS, vegan package v2.6.4). Microbial taxa at the Order-level and CAZyme identity were
320 subsequently fitted to their respective NMDS graphs using the envfit function in the vegan package
321 (v2.6.4) to variables influencing each sample or MAG.
322
323 *2.8 Statistics and Plot Generation.* Statistical analysis and plot generation were performed using
324 R v4.3.1 and RStudio (Posit team, 2023; R Core Team, 2023). Plotting relied on a mixture of
325 functions from the tidyverse collection of R packages (v2.0.0), ggthemes (v4.2.4), ggrepel (v0.9.3),
326 and ggpubr (v0.6.0) (Arnold, 2021; Kassambara, 2023; Slowikowski, 2023; Wickham et al., 2019).
327 Kruskal-Wallis tests were used to evaluate variation across multiple groups (e.g., whether

differences existed across all rivers) using the kruskal.test function, and Mann Whitney U tests were used to evaluate pairwise variation using the wilcox.test function. Spearman correlations were performed using the rcorr function from the Hmisc package (v5.1.0) (Harrell, 2023).

All code and relevant files can be found on ESS-DIVE at https://data.ess-dive.lbl.gov/view/doi:10.15485/2319037 (Danczak et al., 2024). All raw data used in this study can be found on ESS-DIVE using the following DOIs: https://data.ess-dive.lbl.gov/view/doi:10.15485/1509695, https://data.ess-dive.lbl.gov/view/doi:10.15485/1577266, https://data.ess-dive.lbl.gov/view/doi:10.15485/1577260, https://data.ess-dive.lbl.gov/view/doi:10.15485/1577265, https://data.ess-dive.lbl.gov/view/doi:10.15485/1577278, https://data.ess-dive.lbl.gov/view/doi:10.15485/1577263, https://data.ess-dive.lbl.gov/view/doi:10.15485/1576995 (Chu et al., 2019; Danczak et al., 2019; Garayburu-Caruso et al., 2019; Goldman et al., 2019; Renteria et al., 2019; Stegen et al., 2019; Wells et al., 2019). Metagenomic data has been uploaded to NCBI and can be accessed via the Bioproject # PRJNA946291.

## 3. Results and Discussion

*3.1 Globally distributed rivers exhibit strong geographic environmental metabolome signatures despite geochemical commonalities in similar surrounding environments.* Geochemical analyses revealed substantial environmental variation across the seven globally distributed rivers. Looking first at the geochemical patterns using a principal component analysis (Supplemental Figure 2),

we see that each river has characteristic geochemical parameter values. Both surface water and pore water from the Erpe River (ER) are characterized by higher NPOC, K, $SO_4$, Na, Cl, and $NO_2$ concentrations, while the surface water of East Fork Poplar Creek (EFPC) has higher Ca, Mg, $NO_3$, Total N, and P concentrations. Given that both systems are considered either currently or historically wastewater contaminated (Loar et al., 2011; Mueller et al., 2021), we would anticipate elevated levels of many geochemical parameters like NPOC and $NO_3$/$NO_2$ while other geochemical parameters like Ca and Mg are likely due to the underlying lithology. The Altamaha River (AR) is characterized by elevated iron concentrations likely due to the dominant sediment mineralogy, lower pH, and higher NPOC (Supplemental Figure 1) (Bhatti et al., 2009; Roebuck et al., 2020; Shi et al., 2001). The four remaining rivers (and the pore water from EFPC) are defined by lower concentrations of most parameters, particularly the HJ Andrews Watershed (HJA) and the Nisqually River (NR) which are a headwater stream and a glacial melt run-off stream, respectively (Curran et al., 2016; Danczak et al., 2021). Examining carbon data for combined pore and surface water samples (Figure 1c), we observe significant differences in NPOC concentrations across the seven rivers resulting in the formation of three broad groups: ER and AR constitute a high NPOC group; the Jordan River (JR), HJA, and NR form a low NPOC group; the Columbia River (CR) and EFPC have intermediate concentrations between these two groups.

Shifting focus to average environmental metabolome characteristics, each of the seven rivers has a distinct organic matter mixture. The average nominal oxidation state of carbon (NOSC), a metric which can track the thermodynamic bioavailability of organic matter, indicated no groups formed among the rivers. Nevertheless, there were significant differences among some pairwise comparisons (e.g., NR was significantly lower than AR), though few differences exist between

water types (e.g., EFPC surface water was significantly different from EFPC pore water) (Supplemental Table 2, Figure 1d). Average double bond equivalents (DBE) and modified aromaticity index ($AI_{Mod}$) were used to evaluate potential structural patterns of organic matter across rivers and water types. Patterns based upon DBE are more apparent with AR, CR, and HJA exhibiting higher values, ER, JR, and NR exhibiting lower values, each for both water types; DBE for EFPC was different across the pore and surface water. No clear groupings across rivers are apparent in $AI_{Mod}$ in either surface water or pore water samples though EFPC, ER, and JR have significantly different average aromaticity across surface and pore water. Overall, the less urbanized, headwater rivers (e.g., NR, JR) have lower NOSC and DBE overall, while the larger rivers with many tributaries (e.g., AR, CR) have higher NOSC and DBE. The two wastewater and one of the headwater rivers (HJA) have NOSC and DBE values intermediate between the two end-member groups. While similar river types (e.g., headwater, wastewater) group together on some metrics, the specific location of each river plays a substantial role in shaping average characteristics. This pattern is also in keeping with data indicating that the surrounding land (often land coverage) plays a differentiating role in DOM composition (Danczak et al., 2023; Roebuck et al., 2020). Moreover, these patterns are not exclusively a function of river NPOC concentration.

River type (e.g., wastewater vs. headwater) is a significant determinant for molecular formula richness (e.g., the number of molecular formulas). Rivers fell into two groups based on molecular formula richness – AR, CR, EFPC, and ER form a high richness group, while HJA, JR, and NR constitute a low richness group and these groups held across both water types (Figure 1d). The patterns in richness closely mirror the NPOC concentrations, with some disagreement; namely, EFPC and the CR have higher richness relative to their NPOC concentrations. Likely, this is in

397 part driven by wastewater for EFPC and the large number of tributaries that feed into the CR
398 Specifically, wastewater is typically high in diverse types of organic matter while the many
399 tributaries of the CRB bring the potential for varied organic matter (e.g., pyrogenic DOM from
400 wildfires, farmland input, urban, etc.) despite comparatively low overall concentrations (Danczak
401 et al., 2023; Vaughn et al., 2023). The low formula richness group consists of samples that were
402 collected either in or close to the headwaters. This data may suggest that larger rivers under
403 agricultural/urban influence (CR and AR) and wastewater-impacted rivers (EFPC and ER) will
404 have higher richness than less urbanized headwater streams (HJA, JR, and NR). Such a pattern is
405 consistent with observations made within the Upper Mississippi River basin whereby forested
406 rivers had lower molecular formula richness than urban and agricultural rivers (Vaughn et al.,
407 2023). These results, in conjunction with previous studies in diverse basins, indicate that
408 incorporating the surrounding land into predictions regarding DOM development is critical.
409 However, the specific interactions between land type and environmental conditions and
410 subsequent impacts on DOM are not clear.
411
412 *3.2 Ecological null modeling of environmental metabolomes reveals geographically distinct forces*
413 *drive the development of DOM in rivers.* Multivariate analyses revealed that the DOM composition
414 was significantly different across rivers and both water types (Figure 2a). The exceptions to this
415 pattern were HJA and NR which had statistically indistinguishable surface and pore water samples.
416 Both systems have low NPOC values and HJA has a high degree of hydrological connectivity
417 between surface and pore water (Ward et al., 2018; Wondzell, 2006; Wondzell et al., 2007).
418 Importantly, the HJA results differ from our previous study which indicated that the pore water
419 and surface water metabolomes were significantly divergent (Danczak et al., 2021). The

contrasting pattern for HJA between the two studies is likely driven by a shift in scale – by providing more chemodiversity from additional rivers in this study, the surface and pore water HJA metabolomes become much more similar in comparison. Similar results are noted in ecological studies and emphasize that scale has important influences over perceived variation in composition (Hernández, 2020; Ladau and Eloe-Fadrosh, 2019; Levin, 1992; Swenson et al., 2006).

The DOM composition of each river was significantly, but weakly, related to the corresponding geochemical profile (Mantel ρ: 0.1985, p-value: 0.001; PROTEST $m_{12}^2$: 0.7379, p-value: 0.001). A redundancy analysis (RDA) elucidated which geochemical parameters explained the cross-river metabolome differences (Figure 2b). Initial forward/reverse variable selection prior to the RDA suggested that 9 parameters (NPOC, Cl, Mg, TN, $NO_3$, F, P, Fe, and pH) sufficiently captured variability across the samples. The RDA revealed that 6 parameters (Cl, Mg, $NO_3$, NPOC, TN, and P) were associated with the samples from wastewater-impacted streams (ER and the surface water of EFPC), while HJA, JR, and NR were associated with higher pH values and lower concentrations of those 6 parameters. As in other analyses, the pore water of EFPC, AR, and CR were intermediate to the other 4 rivers, though there was an iron signal associated with AR and a weak fluoride signal for CR. These relationships indicate a cohesion between geochemical and organic matter variability and suggest that geochemical condition might be a stronger predictor than geographical location (e.g., NR and JR are more similar each other than other geographically closer locations). While useful, these analyses do not tell us how the geochemistry is related to the ecological processes which shape the environmental metabolome.

Following protocols established via meta-metabolome ecology, we observed that rivers experienced significantly different ecological assembly processes (Danczak et al., 2020). Briefly, this method leverages a null model called β-nearest taxon index (βNTI) to identify ecological processes impacting the environmental metabolome. Based upon the βNTI value, we can assign either a deterministic (e.g., selection; $|\beta NTI| > 2$) or stochastic (e.g., dispersal; $|\beta NTI| < 2$) influence. Deterministic processes can be further separated based upon their sign. When $\beta NTI < -2$, the communities were assembled due to homogenous selection and are more similar than would be expected by random chance. When $\beta NTI > 2$, communities experienced variable selection because they are more divergent than would be expected by random chance.

Comparing the environmental metabolomes among rivers, each river experiences variable selection overall ($\beta NTI > 2$), with ER having the strongest variable selection and AR having the weakest, bordering on stochastic assembly (Figure 3a). By relating βNTI values to geochemical parameters, we can reveal which parameters are important in structuring environmental metabolomes (Danczak et al., 2021, 2016; Stegen et al., 2016b). This analysis revealed that 8 parameters (Ca, $SO_4$, K, Mg, NPOC, Na, $NO_2$, and TN in order of correlation strength) were significantly positively related to global βNTI values and iron was negatively related (Supplemental Table 3). Looking to those parameters which were influential in the RDA (Figure 2b), we see overlap between these relationships suggesting that many of the influential parameters may impact ecological processes likely through either metabolic or abiotic interaction. For example, some parameters like calcium, potassium, or sodium might be indicative of underlying lithology (e.g., calcium carbonate content) or anthropogenic influence and suggest that certain geological characteristics might lead to deterministic, variable selection (e.g., larger differences in

466 sodium are related to higher βNTI). Alternatively, other metrics like sulfate, total nitrogen, or
467 nitrite may highlight specific nutrient, metabolic, or respiratory activity related to divergent
468 environmental metabolomes. Iron being the only negative relationship may point to a few
469 possibilities: 1) iron concentrations fluctuate inversely with one of the positively related
470 geochemical variables (e.g., pH), or 2) higher iron concentrations are associated with
471 homogeneous selection in environmental metabolomes via metabolic (e.g., iron reduction) or
472 abiotic (e.g., organometallic) interactions (Lv et al., 2016). Iron's interpretation is further
473 complicated in that many sites had iron concentrations that were below detection (Supplemental
474 Figure 1). Regardless of the specific interaction, geochemical conditions are significantly related
475 to the ongoing ecological processes shaping environmental metabolomes across seven
476 geographically distinct rivers.
477
478 By analyzing βNTI comparisons within a given group (here within rivers), we can evaluate local,
479 rather than global, processes (Figure 3b). EFPC and ER are structured by variable selection (βNTI
480 > 2), HJA and JR are structured by homogeneous selection (βNTI < -2), and AR, CR, and NR are
481 structured by stochastic processes (2 > βNTI > -2). Firstly, we see that the wastewater rivers in
482 this study are governed by distinct ecological assembly processes. Secondly, rivers with low
483 carbon content were more impacted by homogeneous processes. These low carbon rivers also have
484 lower geochemical variability as indicated by principal component and the redundancy analyses.
485 Combined, this suggests that low environmental variability may lead to homogenized organic
486 matter assemblages. Finally, we see stochastic (i.e., dispersal-based) processes impacting rivers
487 regardless of geochemical condition (e.g., low NPOC, low iron NR and high NPOC, high iron AR
488 were impacted by stochastic assembly). These results point to an observation whereby a priori

river type appears to be an important characteristic of DOM development in some cases (e.g., ER as wastewater, HJA as headwaters) while dispersal may dominate in others. Similar patterns have been discussed previously whereby land use has been related to DOM diversity (Danczak et al., 2023; Roebuck et al., 2020; Vaughn et al., 2023). Results from these studies support our ecological observations, and indicate that geochemical conditions and the specific location of a river (including the historical influences up to the point in time when sampling occurred) are important in understanding the development of DOM.

*3.3 The global evolutionary history of river corridor microbes is more significant in dictating carbon degradation potential than local pressures.* Microbes are key drivers of biogeochemical cycles and are linked to river corridor organic matter. By understanding the putative functions of resident microbes, we can better understand the forces driving DOM development (Graham et al., 2018, 2016). To evaluate the functional potential as it relates to organic matter diversity, a subset of surface water and pore water samples from AR, CR, EFPC, ER, and JR (1 from each, 10 total) were analyzed using shotgun metagenomic sequencing and a total of 302 metagenomic assembled genomes (MAGs) of medium or higher quality were obtained. These MAGs belong to diverse lineages which are detected across all river corridors (Figure 4a). At higher taxonomic levels, members of the FCB Group and PVC Superphylum were broadly distributed. At lower taxonomic levels, MAGs from the Order Burkholderiales represent at least 1% relative coverage in all rivers aside from JR (Supplemental Figure 3). These Burkholderiales MAGs have consistent carbon, nitrogen, and sulfur processing functional potential, though those found in ER often encoded both an acetate kinase and pyruvate formate lyase and more diverse nitrogen metabolisms (Supplemental Figure 4). This may indicate that these organisms encounter more diverse carbon

512 sources (e.g., short chain fatty acids) and nitrogen types due to their wastewater habitat (Cao et al.,
513 2013; Zhou et al., 2023). While no Burkholderiales were observed within JR, this river shared
514 members of the Order Flavobacteriales with AR, CR, and EFPC and the Order Rhizobiales with
515 AR. The core functional potential of these MAGs is also consistent across the sampled river
516 corridors (Supplemental Figure 4). However, sequences annotated as PEP carboxylase, an enzyme
517 involved in carbon fixation, are present in the Rhizobiales and Flavobacteriales in JR only.
518
519 A similar consistency in core metabolism is observed when examining less broadly distributed
520 taxa (Figure 4b; Supplemental Figure 5). Most Orders (aside from UBA9983) encode common
521 carbon, nitrogen, and sulfur metabolisms with limited variation. The 3 Campylobacterales MAGs
522 found in ER encode pyruvate:ferrodoxin oxidoreductase (PFOR), pyruvate water dikinase, formate
523 dehydrogenase. The Chitinophagales were observed both in CR (4 MAGs) and JR (2 MAGs), but
524 exhibited geographical variations: namely, the JR variants encoded PFOR, the RNF complex,
525 formate-tetrahydrofolate ligase, and a Na-transporting NADH:ubiquinone oxidoreductase. A
526 single Methanomicrobiales MAG was recovered from EFPC and encoded the potential for carbon
527 fixation via the Wood-Ljungdahl Pathway (CO-methylated acetyl-CoA synthase) and for
528 methanogenesis via $CO_2$. Many Nanopelagicales MAGs were primarily detected in AR (19
529 MAGs), though some were observed in CR (1 MAGs) and ER (3 MAGs) with low coverage. These
530 genomes were characterized by the presence of genes encoding an L-lactate dehydrogenase, a
531 partial malonate semialdehyde pathway, a PEP carboxylase, formate-tetrahydrofolate ligase, and
532 3-phenylpropionate dioxygenase. Finally, the UBA9983 completely lacks genes belonging to the
533 Citric Acid Cycle and has limited metabolic versatility in keeping with other members of the
534 Patescibacteria (also known as the Candidate Phyla Radiation) (Castelle et al., 2017; Vigneron et

535 al., 2023). UBA9983 also exhibits geographical patterns: the 4 MAGs detected in AR encode
536 pyruvate water dikinase and D-lactate dehydrogenases while the 15 ER MAGs encode fewer
537 SCFA enzymes and more diverse nitrogen cycling genes.
538
539 These patterns lead to two key observations for these microbial communities: 1) despite significant
540 geographical distances and phylogenetic variation, MAGs observed across the five river corridors
541 have largely similar core metabolic functions, and 2) differences that do arise appear to be related
542 to carbon-type utilization. This type of conservation in functional potential despite phylogenetic
543 variation has been observed in other ecosystems and likely indicates that microbial communities
544 across these five river corridors face many common pressures, though not necessarily dispersal
545 events (Isobe et al., 2020; Martiny et al., 2013; Walkup et al., 2023). In turn, this leads to the
546 development of a common set of functions encoded by varied taxa. However, while many
547 pressures may be similar, each river corridor has a set of unique conditions as evidenced by the
548 geochemical and organic C measurements above (Figure 1; Supplemental Figure 1). This gives
549 rise to a mixture of shared core metabolism and unique peripheral metabolisms, sometimes within
550 the same taxa. For example, while Rhizobiales and Flavobacterales are found in 3 and 4 river
551 corridors, respectively, only those in JR contain a gene characteristic of carbon fixation. The JR
552 also featured some of the lowest NPOC concentrations observed within the dataset. Taken
553 together, this suggests that conditions in the JR might select for carbon fixation due to low carbon
554 concentration. While this is only one example, the other variations in functional potential detected
555 across rivers suggest a more differentiating impact of carbon conditions than other variables (e.g.,
556 the differential presence of PFOR, formate-active sequences, etc.).
557

558 Diving more deeply into the carbon cycling potential of the Orders described above revealed
559 potential taxonomically conserved niche differentiation across the seven rivers (Figure 5a).
560 Specifically, we see that the Verrucomicrobiales and Rhizobiales (both present in JR) contain more
561 diverse and significantly divergent CAZymes than the UBA9983 (characteristic of ER). An
562 expanded CAZyme analysis for all MAGs revealed that potential niche differentiation exists across
563 the taxonomic groups found in the five river corridors. This is captured by the high cross-
564 taxonomic CAZyme beta-diversity (e.g., distance in the ordination) (Figure 5b). For example,
565 while we see that the Burkholderiales and Nanopelagicales have comparable CAZyme alpha
566 diversity, the distance between them in the ordination is high suggesting divergent roles in
567 environmental carbon degradation. Meanwhile, phylogenetic conservation within taxonomic
568 groups is also suggested by the relatively low within-group beta-diversity. For example, we see
569 that MAGs from the Burkholderiales cluster closely together despite being obtained from 4
570 geographically distinct rivers (AR, CR, ER, and EFPC). These patterns of taxonomic niche
571 differentiation and phylogenetic conservation of functional potential suggest that the global
572 evolutionary history is more important to an organism's CAZyme pool than local evolutionary
573 pressures. When combined with the above observation that many differences arise in the functional
574 potential related to smaller carbon compounds, clades are selected for based upon their full carbon
575 degradation potential which can be common across rivers (e.g., Burkholderiales) or specific to an
576 individual river (e.g., UBA9983).
577
578 When various microbial diversity measurements (i.e., Shannon diversity for the overall
579 community, for all CAZYmes contained in the genomes of the community, and for only glycosyl
580 hydrolases in those genomes) are related to DOM richness (i.e., the number of molecular formulas

581 within a sample), we observe significant correlations (Figure 5c). Specifically, the diversity of
582 genes putatively involved in carbon cycling (e.g., CAZymes, glycosyl hydrolases) have slightly
583 stronger relationships with DOM richness than overall diversity (Shannon's $\rho = 0.75$ vs. CAZyme
584 and GH $\rho = 0.82 – 0.83$). This highlights a connection between DOM variety within an ecosystem
585 and the taxonomic and functional potential of microorganisms. When considered in concert with
586 the potential niche partitioning described before, these results suggest that microorganisms may
587 assemble into a community based on the exposure to varied types of DOM, at least in part. Given
588 that carbon cycling potential evolved prior to the dissemination of specific microbial taxa as
589 suggested by the conservation of degradation capability within a given taxonomy (Figure 5ab),
590 DOM might act as a filter for selecting for present groups and may point to causality in the
591 composition of microbial communities. We suggest caution, however, in extrapolating these
592 results too broadly as they were obtained from a limited dataset, from limited data types.
593 Exploration of a greater number of samples across a wider range of biomes while leveraging varied
594 technologies (e.g., metatranscriptomics, metaproteomics, lipidomics, etc.) are necessary before
595 suggesting anything authoritative. Despite these limitations, these results offer a promising
596 glimpse into the ecological interconnections between microbes and the DOM they consume.
597

598 **4. Conclusions**

599 The combination of meta-metabolome ecology, in-depth geochemical analyses, and shotgun
600 metagenomics provided insights into the potential driving forces behind organic matter chemistry
601 and the interconnections between microbial communities and organic matter across seven rivers.
602 The geochemistry of an environment is significantly related to metabolome composition as
603 suggested by traditional multivariate analyses (e.g., Mantel and permuted PROCRUSTES tests)

604 and null modeling approaches. This pattern was expected due to the distinct geochemical
605 signatures of each location. However, our null modeling revealed that there are certain key
606 geochemical parameters that may play a role in structuring organic matter. Calcium, NPOC,
607 sulfate, potassium, sodium, magnesium, total nitrogen, and nitrate consistently were detected as
608 important predictors for variable selection (e.g., positively related to βNTI) (Supplemental Table
609 4). These geochemical characteristics largely reflect watershed characteristics (e.g. soil
610 characteristics, land cover) and geological processes (e.g., underlying lithology). For example,
611 calcium concentrations in circumneutral conditions are largely controlled by natural weathering
612 processes (Weyhenmeyer et al., 2019). Relationships with calcium (or other primarily weathering
613 derived characteristics) might provide insight into organic matter dynamics that are associated with
614 certain geological histories. They might also point to specific land-use or metabolic processes
615 driving organic matter compositions. Land-use and/or land type (i.e., forested, urban, wastewater,
616 etc.) significantly impacts carbon flux (Regnier et al., 2013) and potentially explains relationships
617 with NPOC. Many of these variables, however, are also involved in numerous metabolic pathways
618 and could point to organismal activity. For example, previous research in HJA and CR revealed
619 that nitrogen metabolism is an important component of DOM processing within those rivers
620 (Danczak et al., 2021; Graham et al., 2018; Stegen et al., 2018). Therefore, the relationships
621 between total nitrogen or nitrate and DOM may represent dynamics in associated nutrient or
622 energy pathways.
623
624 Relationships within the microbial communities were more challenging to identify due to the
625 limited sample number, but a few clear patterns emerged. First, we observed members of the
626 Burkholderiales in 4 of the 5 analysed rivers (AR, CR, EFPC, and ER) and members of the

627 Flavobacteriales in 3 of the 5 rivers (AR, CR, and JR). Focusing on carbon cycling potential, it
628 appears that the Burkholderiales play similar roles in each river despite significant geographical
629 distances (Supplementary Figure 3). Second, those lineages which are only found in fewer rivers
630 appear to play diversified roles, potentially related to environmental conditions of a given area.
631 For example, members of the Patescibacteria (UBA9983) are significantly functionally divergent
632 from the Verrucomicrobiales and Rhizobiales. Whereas the Verrucomicrobiales and Rhizobiales
633 appear to have similar carbon cycling potentials as many aerobic organisms, the UBA9983 MAGs
634 encode no respiratory capabilities, have reduced CAZyme diversity, and seem to be obligate
635 fermenters consistent with their taxonomic assignment (Castelle et al., 2017; Vigneron et al.,
636 2023). These differences in putative niche space represent the potential interactions between the
637 microbial community and DOM. While members with a broader potential niche space, like the
638 Rhizobiales or Flavobacteriales, may be able to access more DOM types due to a greater diversity
639 of encoded CAZymes, members with narrower potential niche space (e.g., UBA9983) would be
640 able to target less DOM. When DOM-based diversity measurements are related to these microbial
641 data, we observed significant interconnections highlighting putative role for DOM in selecting for
642 subsets of microbial taxa.
643
644 By examining organic matter chemistry in conjunction with geochemistry and microbial functional
645 potential, we were able to identify general trends across seven globally distributed rivers. Namely,
646 we observed that both geochemical conditions and the specific location of a river (i.e., the type of
647 the river, the land use history of the surrounding area, etc.) are important in understanding riverine
648 DOM composition. Additionally, we observed that these same factors significantly influenced the
649 underlying ecological assembly processes structuring DOM, often in a scale-dependent manner.

Finally, our metagenomic analyses indicated that members within the microbial communities are differentiated by their carbon degradation potential with molecular diversity serving as a putative selective filter. Together, these results give us a glimpse into the interconnections between DOM development and microbial functional potential while also highlighting environmental parameters that might be broadly important in understanding DOM in rivers.

**References**


Arnold, J.B., 2021. ggthemes: Extra Themes, Scales and Geoms for "ggplot2."

Arntzen, E.V., Geist, D.R., Dresel, P.E., 2006. Effects of fluctuating river flow on groundwater/surface water mixing in the hyporheic zone of a regulated, large cobble bed river. River Res. Appl. 22, 937–946. https://doi.org/10.1002/rra.947

Bhatti, A.M., Rundquist, D., Schalles, J., Ramirez, L., Nasu, S., 2009. A comparison between above-water surface and subsurface spectral reflectances collected over inland waters. Geocarto Int. 24, 133–141. https://doi.org/10.1080/10106040802460707

Bramer, L.M., White, A., 2019. ftmsRanalysis: Analysis and visualization tools for FT-MS data. R package version 1.0.0. [WWW Document]. URL https://github.com/EMSL-Computing/ftmsRanalysis

Bramer, L.M., White, A.M., Stratton, K.G., Thompson, A.M., Claborne, D., Hofmockel, K., McCue, L.A., 2020. FtmsRanalysis: An R package for exploratory data analysis and interactive visualization of FT-MS data. PLoS Comput. Biol. 16, 1–12. https://doi.org/10.1371/journal.pcbi.1007654

Brooks, S.C., Southworth, G.R., 2011. History of mercury use and environmental contamination at the Oak Ridge Y-12 Plant. Environ. Pollut. 159, 219–228.

https://doi.org/10.1016/j.envpol.2010.09.009

Buser-Young, J.Z., Garcia, P.E., Schrenk, M.O., Regier, P.J., Ward, N.D., Biçe, K., Brooks, S.C., Freeman, E.C., Lønborg, C., 2023. Determining the biogeochemical transformations of organic matter composition in rivers using molecular signatures. Front. Water 5, 1005792. https://doi.org/10.3389/frwa.2023.1005792

Bushnell, B., 2020. BBTools software package.

Cao, S., Wang, S., Peng, Y., Wu, C., Du, R., Gong, L., Ma, B., 2013. Achieving partial denitrification with sludge fermentation liquid as carbon source: The effect of seeding sludge. Bioresour. Technol. 149, 570–574. https://doi.org/10.1016/j.biortech.2013.09.072

Capella-Gutiérrez, S., Silla-Martínez, J.M., Gabaldón, T., 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics 25, 1972–1973. https://doi.org/10.1093/bioinformatics/btp348

Castelle, C.J., Brown, C.T., Thomas, B.C., Williams, K.H., Banfield, J.F., 2017. Unusual respiratory capacity and nitrogen metabolism in a Parcubacterium (OD1) of the Candidate Phyla Radiation. Sci. Rep. 7, 40101. https://doi.org/10.1038/srep40101

Chaumeil, P.-A., Mussig, A.J., Hugenholtz, P., Parks, D.H., 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics 36, 1925–1927. https://doi.org/10.1093/bioinformatics/btz848

Chu, R.K., Goldman, A.E., Brooks, S.C., Danczak, R.E., Garayburu-Caruso, V.A., Graham, E.B., Jones, M., Jones, N., Lin, X., Morad, J.W., Ren, H., Renteria, L., Resch, C.T., Tfaily, M., Tolic, N., Toyoda, J.G., Wells, J.R., Stegen, J.C., 2019. WHONDRS 48 Hour Diel Cycling Study at the East Fork Poplar Creek in Tennessee, USA. https://doi.org/10.15485/1577278

Cole, J.J., Prairie, Y.T., Caraco, N.F., McDowell, W.H., Tranvik, L.J., Striegl, R.G., Duarte, C.M., Kortelainen, P., Downing, J.A., Middelburg, J.J., Melack, J., 2007. Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget. Ecosystems 10, 171–184. https://doi.org/10.1007/s10021-006-9013-8

Curran, C.A., Grossman, E.E., Magirl, C.S., Foreman, J.R., 2016. Suspended Sediment Delivery to Puget Sound from the Lower Nisqually River, Western Washington, July 2010–November 2011. US Geol. Surv. Sci. Investig. Rep. 2016-5062 17. http://dx.doi.org/10.3133/sir20165062

Danczak, R.E., Chu, R.K., Fansler, S.J., Goldman, A.E., Graham, E.B., Tfaily, M.M., Toyoda, J., Stegen, J.C., 2020. Using metacommunity ecology to understand environmental metabolomes. Nat. Commun. 11, 6369. https://doi.org/10.1038/s41467-020-19989-y

Danczak, R.E., Garayburu-Caruso, V.A., Renteria, L., McKever, S.A., Otenburg, O.C., Grieger, S.R., Son, K., Kaufman, M.H., Fulton, S.G., Roebuck, J.A., Myers-Pigg, A.N., Stegen, J.C., 2023. Riverine organic matter functional diversity increases with catchment size. Front. Water 5, 1087108. https://doi.org/10.3389/frwa.2023.1087108

Danczak, R.E., Goldman, A.E., Chu, R.K., Garayburu-Caruso, V.A., Graham, E.B., He, X., Lin, X., Meile, C., Morad, J.W., Ren, H., Renteria, L., Resch, C.T., Rooze, J., Schalles, J., Tfaily, M., Thomle, J., Tolic, N., Toyoda, J.G., Wells, J.R., Stegen, J.C., 2019. WHONDRS 48 Hour Diel Cycling Study at the Altamaha River in Georgia, USA. https://doi.org/10.15485/1577263

Danczak, R.E., Goldman, A.E., Chu, R.K., Toyoda, J.G., Garayburu-Caruso, V.A., Tolić, N., Graham, E.B., Morad, J.W., Renteria, L., Wells, J.R., Herzog, S.P., Ward, A.S., Stegen, J.C., 2021. Ecological theory applied to environmental metabolomes reveals compositional

divergence despite conserved molecular properties. Sci. Total Environ. 788, 147409. https://doi.org/10.1016/j.scitotenv.2021.147409

Danczak, R.E., Sawyer, A.H., Williams, K.H., Stegen, J.C., Hobson, C., Wilkins, M.J., 2016. Seasonal hyporheic dynamics control coupled microbiology and geochemistry in Colorado River sediments. J. Geophys. Res. Biogeosciences 121, 2976–2987. https://doi.org/10.1002/2016JG003527

D'Andrilli, J., Cooper, W.T., Foreman, C.M., Marshall, A.G., 2015. An ultrahigh-resolution mass spectrometry index to estimate natural organic matter lability. Rapid Commun. Mass Spectrom. 29, 2385–2401. https://doi.org/10.1002/rcm.7400

Danecek, P., Bonfield, J.K., Liddle, J., Marshall, J., Ohan, V., Pollard, M.O., Whitwham, A., Keane, T., McCarthy, S.A., Davies, R.M., Li, H., 2021. Twelve years of SAMtools and BCFtools. GigaScience 10, giab008. https://doi.org/10.1093/gigascience/giab008

Demers, J.D., Blum, J.D., Brooks, S.C., Donovan, P.M., Riscassi, A.L., Miller, C.L., Zheng, W., Gu, B., 2018. Hg isotopes reveal in-stream processing and legacy inputs in East Fork Poplar Creek, Oak Ridge, Tennessee, USA. Environ. Sci. Process. Impacts 20, 686–707. https://doi.org/10.1039/C7EM00538E

Drake, T.W., Raymond, P.A., Spencer, R.G.M., 2018. Terrestrial carbon inputs to inland waters: A current synthesis of estimates and uncertainty. Limnol. Oceanogr. Lett. 3, 132–142. https://doi.org/10.1002/lol2.10055

Edgar, R.C., 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 32, 1792–1797. https://doi.org/10.1093/nar/gkh340

Finn, R.D., Coggill, P., Eberhardt, R.Y., Eddy, S.R., Mistry, J., Mitchell, A.L., Potter, S.C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G.A., Tate, J., Bateman, A., 2016.

The Pfam protein families database: towards a more sustainable future. Nucleic Acids Res. 44, D279–D285. https://doi.org/10.1093/nar/gkv1344

Fudyma, J.D., Chu, R.K., Graf Grachet, N., Stegen, J.C., Tfaily, M.M., 2021. Coupled Biotic-Abiotic Processes Control Biogeochemical Cycling of Dissolved Organic Matter in the Columbia River Hyporheic Zone. Front. Water 2, 1–20. https://doi.org/10.3389/frwa.2020.574692

Garayburu-Caruso, V.A., Goldman, A.E., Chu, R.K., Danczak, R.E., Graham, E.B., Lin, X., Morad, J.W., Ren, H., Renteria, L., Resch, C.T., Tfaily, M., Tolic, N., Toyoda, J.G., Wells, J.R., Stegen, J.C., 2019. WHONDRS 48 Hour Diel Cycling Study at the Columbia River in Washington, USA. https://doi.org/10.15485/1577265

Garayburu-Caruso, V.A., Stegen, J.C., Song, H.-S., Renteria, L., Wells, J., Garcia, W., Resch, C.T., Goldman, A.E., Chu, R.K., Toyoda, J., Graham, E.B., 2020. Carbon Limitation Leads to Thermodynamic Regulation of Aerobic Metabolism. Environ. Sci. Technol. Lett. 7, 517–524. https://doi.org/10.1021/acs.estlett.0c00258

Goldman, A.E., Arnon, S., Bar-Zeev, E., Chu, R.K., Danczak, R.E., Garayburu-Caruso, V.A., Graham, E.B., Lin, X., Morad, J.W., Ren, H., Renteria, L., Resch, C.T., Tfaily, M., Tolic, N., Toyoda, J.G., Wells, J.R., Stegen, J.C., 2019. WHONDRS 48 Hour Diel Cycling Study at the Jordan River, Israel. https://doi.org/10.15485/1577266

Gómez-Gener, L., Rocher-Ros, G., Battin, T., Cohen, M.J., Dalmagro, H.J., Dinsmore, K.J., Drake, T.W., Duvert, C., Enrich-Prast, A., Horgby, Å., Johnson, M.S., Kirk, L., Machado-Silva, F., Marzolf, N.S., McDowell, M.J., McDowell, W.H., Miettinen, H., Ojala, A.K., Peter, H., Pumpanen, J., Ran, L., Riveros-Iregui, D.A., Santos, I.R., Six, J., Stanley, E.H., Wallin, M.B., White, S.A., Sponseller, R.A., 2021. Global carbon dioxide efflux from rivers

enhanced by high nocturnal emissions. Nat. Geosci. 14, 289–294. https://doi.org/10.1038/s41561-021-00722-3

Graham, E.B., Crump, A.R., Kennedy, D.W., Arntzen, E., Fansler, S., Purvine, S.O., Nicora, C.D., Nelson, W., Tfaily, M.M., Stegen, J.C., 2018. Multi 'omics comparison reveals metabolome biochemistry, not microbiome composition or gene expression, corresponds to elevated biogeochemical function in the hyporheic zone. Sci. Total Environ. 642, 742–753. https://doi.org/10.1016/j.scitotenv.2018.05.256

Graham, E.B., Knelman, J.E., Schindlbacher, A., Siciliano, S., Breulmann, M., Yannarell, A., Beman, J.M., Abell, G., Philippot, L., Prosser, J., Foulquier, A., Yuste, J.C., Glanville, H.C., Jones, D.L., Angel, R., Salminen, J., Newton, R.J., Bürgmann, H., Ingram, L.J., Hamer, U., Siljanen, H.M.P., Peltoniemi, K., Potthast, K., Bañeras, L., Hartmann, M., Banerjee, S., Yu, R.Q., Nogaro, G., Richter, A., Koranda, M., Castle, S.C., Goberna, M., Song, B., Chatterjee, A., Nunes, O.C., Lopes, A.R., Cao, Y., Kaisermann, A., Hallin, S., Strickland, M.S., Garcia-Pausas, J., Barba, J., Kang, H., Isobe, K., Papaspyrou, S., Pastorelli, R., Lagomarsino, A., Lindström, E.S., Basiliko, N., Nemergut, D.R., 2016. Microbes as engines of ecosystem function: When does community structure enhance predictions of ecosystem processes? Front. Microbiol. 7, 1–10. https://doi.org/10.3389/fmicb.2016.00214

Grunert, B.K., Tzortziou, M., Neale, P., Menendez, A., Hernes, P., 2021. DOM degradation by light and microbes along the Yukon River-coastal ocean continuum. Sci. Rep. 11, 10236. https://doi.org/10.1038/s41598-021-89327-9

Harrell, F.E., 2023. Hmisc: Harrell Miscellaneous.

Hernández, F., 2020. Ecological Discord and the Importance of Scale in Scientific Inquiry. J.

Wildl. Manag. 84, 1427–1434. https://doi.org/10.1002/jwmg.21942

Hu, A., Jang, K., Meng, F., Stegen, J., Tanentzap, A.J., Choi, M., Lennon, J.T., Soininen, J., Wang, J., 2022. Microbial and Environmental Processes Shape the Link between Organic Matter Functional Traits and Composition. Environ. Sci. Technol. 56, 10504–10516. https://doi.org/10.1021/acs.est.2c01432

Hughey, C.A., Hendrickson, C.L., Rodgers, R.P., Marshall, A.G., Qian, K., 2001. Kendrick Mass Defect Spectrum: A Compact Visual Analysis for Ultrahigh-Resolution Broadband Mass Spectra. Anal. Chem. 73, 4676–4681. https://doi.org/10.1021/ac010560w

Isobe, K., Bouskill, N.J., Brodie, E.L., Sudderth, E.A., Martiny, J.B.H., 2020. Phylogenetic conservation of soil bacterial responses to simulated global changes. Philos. Trans. R. Soc. B Biol. Sci. 375, 20190242. https://doi.org/10.1098/rstb.2019.0242

Joshi, N., Fass, J., 2011. Sickle: A sliding-window, adaptive, quality-based trimming tool for FastQ files.

Kaiser, K., Canedo-Oropeza, M., McMahon, R., Amon, R.M.W., 2017. Origins and transformations of dissolved organic matter in large Arctic rivers. Sci. Rep. 7, 13064. https://doi.org/10.1038/s41598-017-12729-1

Kanehisa, M., Sato, Y., Kawashima, M., Furumichi, M., Tanabe, M., 2016. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 44, D457–D462. https://doi.org/10.1093/nar/gkv1070

Kang, D.D., Froula, J., Egan, R., Wang, Z., 2015. MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. PeerJ 3, e1165. https://doi.org/10.7717/peerj.1165

Kassambara, A., 2023. ggpubr: “ggplot2” Based Publication Ready Plots.

Koch, B.P., Dittmar, T., 2006. From mass to structure: an aromaticity index for high-resolution mass data of natural organic matter. Rapid Commun. Mass Spectrom. 20, 926–932. https://doi.org/10.1002/rcm.2386

Kujawinski, E.B., Behn, M.D., 2006. Automated analysis of electrospray ionization fourier transform ion cyclotron resonance mass spectra of natural organic matter. Anal. Chem. 78, 4363–4373. https://doi.org/10.1021/ac0600306

Ladau, J., Eloe-Fadrosh, E.A., 2019. Spatial, Temporal, and Phylogenetic Scales of Microbial Ecology. Trends Microbiol. 27, 662–669. https://doi.org/10.1016/j.tim.2019.03.003

LaRowe, D.E., Van Cappellen, P., 2011. Degradation of natural organic matter: A thermodynamic analysis. Geochim. Cosmochim. Acta 75, 2030–2042. https://doi.org/10.1016/j.gca.2011.01.020

Levin, S.A., 1992. The Problem of Pattern and Scale in Ecology: The Robert H. MacArthur Award Lecture. Ecology 73, 1943–1967. https://doi.org/10.2307/1941447

Li, D., Liu, C.-M., Luo, R., Sadakane, K., Lam, T.-W., 2015. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. Bioinformatics 31, 1674–1676. https://doi.org/10.1093/bioinformatics/btv033

Liu, Q., Jiang, Y., Tian, Y., Hou, Z., He, K., Fu, L., Xu, H., 2019. Impact of land use on the DOM composition in different seasons in a subtropical river flowing through a region undergoing rapid urbanization. J. Clean. Prod. 212, 1224–1231. https://doi.org/10.1016/j.jclepro.2018.12.030

Loar, J.M., Stewart, A.J., Smith, J.G., 2011. Twenty-five years of ecological recovery of east fork poplar creek: Review of environmental problems and remedial actions. Environ. Manage. 47, 1010–1020. https://doi.org/10.1007/s00267-011-9625-4

Lv, J., Zhang, S., Wang, S., Luo, L., Cao, D., Christie, P., 2016. Molecular-Scale Investigation with ESI-FT-ICR-MS on Fractionation of Dissolved Organic Matter Induced by Adsorption on Iron Oxyhydroxides. Environ. Sci. Technol. 50, 2328–2336. https://doi.org/10.1021/acs.est.5b04996

Martiny, A.C., Treseder, K., Pusch, G., 2013. Phylogenetic conservatism of functional traits in microorganisms. ISME J. 7, 830–838. https://doi.org/10.1038/ismej.2012.160

Minor, E.C., Steinbring, C.J., Longnecker, K., Kujawinski, E.B., 2012. Characterization of dissolved organic matter in Lake Superior and its watershed using ultrahigh resolution mass spectrometry. Org. Geochem. 43, 1–11. https://doi.org/10.1016/j.orggeochem.2011.11.007

Mueller, B.M., Schulz, H., Danczak, R.E., Putschew, A., Lewandowski, J., 2021. Simultaneous attenuation of trace organics and change in organic matter composition in the hyporheic zone of urban streams. Sci. Rep. 11, 4179. https://doi.org/10.1038/s41598-021-83750-8

Nurk, S., Meleshko, D., Korobeynikov, A., Pevzner, P.A., 2017. metaSPAdes: a new versatile metagenomic assembler. Genome Res. 27, 824–834. https://doi.org/10.1101/gr.213959.116

Oksanen, J., Blanchet, F.G., Friendly, M., Kindt, R., Legendre, P., McGlinn, D., Minchin, P.R., O'Hara, R.B., Simpson, G.L., Solymos, P., Stevens, M.H.H., Szoecs, E., Wagner, H., 2019. vegan: Community Ecology Package [WWW Document]. URL https://cran.r-project.org/package=vegan

Olm, M.R., Brown, C.T., Brooks, B., Banfield, J.F., 2017. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. ISME J. 11, 2864–2868. https://doi.org/10.1038/ismej.2017.126

Osterholz, H., Singer, G., Wemheuer, B., Daniel, R., Simon, M., Niggemann, J., Dittmar, T., 2016. Deciphering associations between dissolved organic molecules and bacterial

857 communities in a pelagic marine system. ISME J. 10, 1717–1730.
858 https://doi.org/10.1038/ismej.2015.231
859 Parks, D.H., Imelfort, M., Skennerton, C.T., Hugenholtz, P., Tyson, G.W., 2015. CheckM:
860 assessing the quality of microbial genomes recovered from isolates, single cells, and
861 metagenomes. Genome Res. 25, 1043–1055. https://doi.org/10.1101/gr.186072.114
862 Peng, Y.Y., Leung, H.C.M., Yiu, S.M., Chin, F.Y.L., 2012. IDBA-UD: a de novo assembler for
863 single-cell and metagenomic sequencing data with highly uneven depth. Bioinformatics 28,
864 1420–1428. https://doi.org/10.1093/bioinformatics/bts174
865 Posit team, 2023. RStudio: Integrated Development Environment for R. Posit Software, PBC,
866 Boston, MA.
867 Price, M.N., Dehal, P.S., Arkin, A.P., 2010. FastTree 2 – Approximately Maximum-Likelihood
868 Trees for Large Alignments. PLoS ONE 5, e9490.
869 https://doi.org/10.1371/journal.pone.0009490
870 R Core Team, 2023. R: A Language and Environment for Statistical Computing. R Foundation
871 for Statistical Computing, Vienna, Austria.
872 Rawlings, N.D., Barrett, A.J., Thomas, P.D., Huang, X., Bateman, A., Finn, R.D., 2018. The
873 MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a
874 comparison with peptidases in the PANTHER database. Nucleic Acids Res. 46, D624–
875 D632. https://doi.org/10.1093/nar/gkx1134
876 Regnier, P., Friedlingstein, P., Ciais, P., Mackenzie, F.T., Gruber, N., Janssens, I.A., Laruelle,
877 G.G., Lauerwald, R., Luyssaert, S., Andersson, A.J., Arndt, S., Arnosti, C., Borges, A.V.,
878 Dale, A.W., Gallego-Sala, A., Goddéris, Y., Goossens, N., Hartmann, J., Heinze, C., Ilyina,
879 T., Joos, F., LaRowe, D.E., Leifeld, J., Meysman, F.J.R., Munhoven, G., Raymond, P.A.,

Spahni, R., Suntharalingam, P., Thullner, M., 2013. Anthropogenic perturbation of the carbon fluxes from land to ocean. Nat. Geosci. 6, 597–607. https://doi.org/10.1038/ngeo1830

Renteria, L., Goldman, A.E., Chu, R.K., Danczak, R.E., Garayburu-Caruso, V.A., Graham, E.B., Lin, X., Morad, J.W., Ren, H., Resch, C.T., Tfaily, M., Tolic, N., Toyoda, J.G., Wells, J.R., Znotinas, K.R., Stegen, J.C., 2019. WHONDRS 48 Hour Diel Cycling Study at the Nisqually River, WA. https://doi.org/10.15485/1576995

Roebuck, J.A., Seidel, M., Dittmar, T., Jaffé, R., 2020. Controls of Land Use and the River Continuum Concept on Dissolved Organic Matter Composition in an Anthropogenically Disturbed Subtropical Watershed. Environ. Sci. Technol. 54, 195–206. https://doi.org/10.1021/acs.est.9b04605

Saarela, T., Zhu, X., Jäntti, H., Ohashi, M., Ide, J., Siljanen, H., Pesonen, A., Aaltonen, H., Ojala, A., Nishimura, H., Kekäläinen, T., Jänis, J., Berninger, F., Pumpanen, J., 2022. Dissolved organic matter composition regulates microbial degradation and carbon dioxide production in pristine subarctic rivers (preprint). Biogeochemistry: Rivers & Streams. https://doi.org/10.5194/bg-2022-225

Shi, W., Sun, M.-Y., Molina, M., Hodson, R.E., 2001. Variability in the distribution of lipid biomarkers and their molecular isotopic composition in Altamaha estuarine sediments: implications for the relative contribution of organic matter from various sources. Org. Geochem. 32, 453–467. https://doi.org/10.1016/S0146-6380(00)00189-3

Slowikowski, K., 2023. ggrepel: Automatically Position Non-Overlapping Text Labels with "ggplot2."

Song, H.-S., Stegen, J.C., Graham, E.B., Lee, J.-Y., Garayburu-Caruso, V.A., Nelson, W.C.,

Chen, X., Moulton, J.D., Scheibe, T.D., 2020. Representing Organic Matter Thermodynamics in Biogeochemical Reactions via Substrate-Explicit Modeling. Front. Microbiol. 11, 531756. https://doi.org/10.3389/fmicb.2020.531756

Stegen, J.C., Fredrickson, J.K., Wilkins, M.J., Konopka, A.E., Nelson, W.C., Arntzen, E.V., Chrisler, W.B., Chu, R.K., Danczak, R.E., Fansler, S.J., Kennedy, D.W., Resch, C.T., Tfaily, M., 2016a. Groundwater–surface water mixing shifts ecological assembly processes and stimulates organic carbon turnover. Nat. Commun. 7, 11237. https://doi.org/10.1038/ncomms11237

Stegen, J.C., Goldman, A.E., 2018. WHONDRS: a Community Resource for Studying Dynamic River Corridors. mSystems 3. https://doi.org/10.1128/mSystems.00151-18

Stegen, J.C., Goldman, A.E., Chu, R.K., Danczak, R.E., Garayburu-Caruso, V.A., Graham, E.B., Lin, X., Morad, J.W., Ren, H., Renteria, L., Resch, C.T., Tfaily, M., Tolic, N., Toyoda, J., Wells, J., Herzog, S., Wondzell, S.M., 2019. WHONDRS 48 Hour Diel Cycling Study at HJ Andrews Experimental Forest Watershed 1 (WS1). https://doi.org/10.15485/1509695

Stegen, J.C., Johnson, T., Fredrickson, J.K., Wilkins, M.J., Konopka, A.E., Nelson, W.C., Arntzen, E.V., Chrisler, W.B., Chu, R.K., Fansler, S.J., Graham, E.B., Kennedy, D.W., Resch, C.T., Tfaily, M., Zachara, J., 2018. Influences of organic carbon speciation on hyporheic corridor biogeochemistry and microbial ecology. Nat. Commun. 9, 585. https://doi.org/10.1038/s41467-018-02922-9

Stegen, J.C., Konopka, A., McKinley, J.P., Murray, C., Lin, X., Miller, M.D., Kennedy, D.W., Miller, E.A., Resch, C.T., Fredrickson, J.K., 2016b. Coupling among Microbial Communities, Biogeochemistry, and Mineralogy across Biogeochemical Facies. Sci. Rep. 6, 1–14. https://doi.org/10.1038/srep30553

Stegen, J.C., Lin, X., Fredrickson, J.K., Chen, X., Kennedy, D.W., Murray, C.J., Rockhold, M.L., Konopka, A., 2013. Quantifying community assembly processes and identifying features that impose them. ISME J. 7, 2069–79. https://doi.org/10.1038/ismej.2013.93

Stegen, J.C., Lin, X., Fredrickson, J.K.J.K., Konopka, A.E., 2015. Estimating and mapping ecological processes influencing microbial community assembly. Front. Microbiol. 6, 1–15. https://doi.org/10.3389/fmicb.2015.00370

Stegen, J.C., Lin, X., Konopka, A.E., Fredrickson, J.K., 2012. Stochastic and deterministic assembly processes in subsurface microbial communities. ISME J. 6, 1653–1664. https://doi.org/10.1038/ismej.2012.22

Swenson, N.G., Enquist, B.J., Pither, J., Thompson, J., Zimmerman, J.K., 2006. The problem and promise of scale dependency in community phylogenetics. Ecology 87, 2418–2424. https://doi.org/10.1890/0012-9658(2006)87[2418:TPAPOS]2.0.CO;2

Takagi, K.K., Hunter, K.S., Cai, W.-J., Joye, S.B., 2017. Agents of change and temporal nutrient dynamics in the Altamaha River Watershed. Ecosphere 8, e01519. https://doi.org/10.1002/ecs2.1519

Tanentzap, A.J., Fitch, A., Orland, C., Emilson, E.J.S., Yakimovich, K.M., Osterholz, H., Dittmar, T., 2019. Chemical and microbial diversity covary in fresh water to influence ecosystem functioning. Proc. Natl. Acad. Sci. U. S. A. 116, 24689–24695. https://doi.org/10.1073/pnas.1904896116

Tfaily, M.M., Chu, R.K., Tolić, N., Roscioli, K.M., Anderton, C.R., Paša-Tolić, L., Robinson, E.W., Hess, N.J., 2015. Advanced solvent based methods for molecular characterization of soil organic matter by high-resolution mass spectrometry. Anal. Chem. 87, 5206–5215. https://doi.org/10.1021/acs.analchem.5b00116

Tfaily, M.M., Chu, R.K., Toyoda, J., Tolić, N., Robinson, E.W., Paša-Tolić, L., Hess, N.J., 2017. Sequential extraction protocol for organic matter from soils and sediments using high resolution mass spectrometry. Anal. Chim. Acta 972, 54–61. https://doi.org/10.1016/j.aca.2017.03.031

Tolić, N., Liu, Y., Liyu, A., Shen, Y., Tfaily, M.M., Kujawinski, E.B., Longnecker, K., Kuo, L.J., Robinson, E.W., Paša-Tolić, L., Hess, N.J., 2017. Formularity: Software for Automated Formula Assignment of Natural and Other Organic Matter from Ultrahigh-Resolution Mass Spectra. Anal. Chem. 89, 12659–12665. https://doi.org/10.1021/acs.analchem.7b03318

Vaughn, D.R., Kellerman, A.M., Wickland, K.P., Striegl, R.G., Podgorski, D.C., Hawkings, J.R., Nienhuis, J.H., Dornblaser, M.M., Stets, E.G., Spencer, R.G.M., 2023. Anthropogenic landcover impacts fluvial dissolved organic matter composition in the Upper Mississippi River Basin. Biogeochemistry 164, 117–141. https://doi.org/10.1007/s10533-021-00852-1

Vigneron, A., Cruaud, P., Guyoneaud, R., Goñi-Urriza, M., 2023. Into the darkness of the microbial dark matter in situ activities through expression profiles of Patescibacteria populations. Front. Microbiol. 13, 1073483. https://doi.org/10.3389/fmicb.2022.1073483

Villa, J.A., Smith, G.J., Ju, Y., Renteria, L., Angle, J.C., Arntzen, E., Harding, S.F., Ren, H., Chen, X., Sawyer, A.H., Graham, E.B., Stegen, J.C., Wrighton, K.C., Bohrer, G., 2020. Methane and nitrous oxide porewater concentrations and surface fluxes of a regulated river. Sci. Total Environ. 715. https://doi.org/10.1016/j.scitotenv.2020.136920

Walkup, J., Dang, C., Mau, R.L., Hayer, M., Schwartz, E., Stone, B.W., Hofmockel, K.S., Koch, B.J., Purcell, A.M., Pett-Ridge, J., Wang, C., Hungate, B.A., Morrissey, E.M., 2023. The predictive power of phylogeny on growth rates in soil bacterial communities. ISME

Commun. 3, 73. https://doi.org/10.1038/s43705-023-00281-1

Ward, A.S., Schmadel, N.M., Wondzell, S.M., 2018. Time-Variable Transit Time Distributions in the Hyporheic Zone of a Headwater Mountain Stream. Water Resour. Res. 54, 2017–2036. https://doi.org/10.1002/2017WR021502

Wells, J.R., Goldman, A.E., Chu, R.K., Danczak, R.E., Garayburu-Caruso, V.A., Graham, E.B., Lewandowski, J., Lin, X., Meinikmann, K., Morad, J.W., Müller, B.M., Ren, H., Renteria, L., Resch, C.T., Schulz, H., Tfaily, M., Tolic, N., Toyoda, J.G., Stegen, J.C., 2019. WHONDRS 48 Hour Diel Cycling Study at the Erpe River, Germany. https://doi.org/10.15485/1577260

Weyhenmeyer, G.A., Hartmann, J., Hessen, D.O., Kopáček, J., Hejzlar, J., Jacquet, S., Hamilton, S.K., Verburg, P., Leach, T.H., Schmid, M., Flaim, G., Nõges, T., Nõges, P., Wentzky, V.C., Rogora, M., Rusak, J.A., Kosten, S., Paterson, A.M., Teubner, K., Higgins, S.N., Lawrence, G., Kangur, K., Kokorite, I., Cerasino, L., Funk, C., Harvey, R., Moatar, F., De Wit, H.A., Zechmeister, T., 2019. Widespread diminishing anthropogenic effects on calcium in freshwaters. Sci. Rep. 9, 10450. https://doi.org/10.1038/s41598-019-46838-w

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L., François, R., Grolemund, G., Hayes, A., Henry, L., Hester, J., Kuhn, M., Pedersen, T., Miller, E., Bache, S., Müller, K., Ooms, J., Robinson, D., Seidel, D., Spinu, V., Takahashi, K., Vaughan, D., Wilke, C., Woo, K., Yutani, H., 2019. Welcome to the Tidyverse. J. Open Source Softw. 4, 1686. https://doi.org/10.21105/joss.01686

Wohl, E., Hall, R.O., Lininger, K.B., Sutfin, N.A., Walters, D.M., 2017. Carbon dynamics of river corridors and the effects of human alterations. Ecol. Monogr. 87, 379–409. https://doi.org/10.1002/ecm.1261

Wondzell, S.M., 2006. Effect of morphology and discharge on hyporheic exchange flows in two small streams in the Cascade Mountains of Oregon, USA. Hydrol. Process. 20, 267–287. https://doi.org/10.1002/hyp.5902

Wondzell, S.M., Gooseff, M.N., McGlynn, B.L., 2007. Flow velocity and the hydrologic behavior of streams during baseflow. Geophys. Res. Lett. 34, L24404. https://doi.org/10.1029/2007GL031256

Woodcroft, B.J., n.d. CoverM.

Xenopoulos, M.A., Barnes, R.T., Boodoo, K.S., Butman, D., Catalán, N., D'Amario, S.C., Fasching, C., Kothawala, D.N., Pisani, O., Solomon, C.T., Spencer, R.G.M., Williams, C.J., Wilson, H.F., 2021. How humans alter dissolved organic matter composition in freshwater: relevance for the Earth's biogeochemistry. Biogeochemistry 154, 323–348. https://doi.org/10.1007/s10533-021-00753-3

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F., Xu, Y., 2012. dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 40, W445–W451. https://doi.org/10.1093/nar/gks479

Yu, G., Smith, D.K., Zhu, H., Guan, Y., Lam, T.T.-Y.Y., 2017. ggtree: an R package for visualization and annotation of phylogenetic trees with their covariates and other associated data. Methods Ecol. Evol. 8, 28–36. https://doi.org/10.1111/2041-210X.12628

Zarnetske, J.P., Bouda, M., Abbott, B.W., Saiers, J., Raymond, P.A., 2018. Generality of Hydrologic Transport Limitation of Watershed Organic Carbon Flux Across Ecoregions of the United States. Geophys. Res. Lett. 45, 11,702-11,711. https://doi.org/10.1029/2018GL080005

Zhou, Y., Zhu, Y., Zhu, J., Li, C., Chen, G., 2023. A Comprehensive Review on Wastewater

Nitrogen Removal and Its Recovery Processes. Int. J. Environ. Res. Public. Health 20, 3429. https://doi.org/10.3390/ijerph20043429

**Acknowledgments**

Many people were involved in the collection of samples for this project, including but not limited to Xiaojia He, Jurjen Rooze, and Katherine Znotinas. Pacific Northwest National Laboratory is operated by Battelle Memorial Institute for the U.S. Department of Energy under Contract No. DE-AC05-76RL01830. This research was supported by the U.S. Department of Energy (DOE), Office of Biological and Environmental Research (BER), as part of BER's Subsurface Biogeochemistry Research Program (SBR). This contribution originates from the SBR Scientific Focus Area (SFA) at the Pacific Northwest National Laboratory (PNNL). A portion of this work was performed at the Environmental Molecular Science Laboratory User Facility.

**Author Contributions**

R.E.D., A.E.G., and J.C.S. conceptualized the study. V.A.G., J.W.M., L.R., J.R.H, M.J., N.J., B.M.M., H.S., J.S., S.A., E.B-Z. collected samples. V.A.G., J.W.M., L.R., J.R.H analyzed anions/cations. R.K.C and J.G.T. collected FTICR-MS data and assisted with analyses. M.A.B. performed bioinformatic analyses. R.E.D. aggregated data, performed ecological and statistical analyses, and analyzed microbial functional potential. R.E.D. drafted the manuscript but all authors contributed to the writing.

**Competing Interests**

The authors declare no competing financial interests.

## Correspondence

Correspondence can be directed at Robert.Danczak@pnnl.gov

## Figure and Table Legends

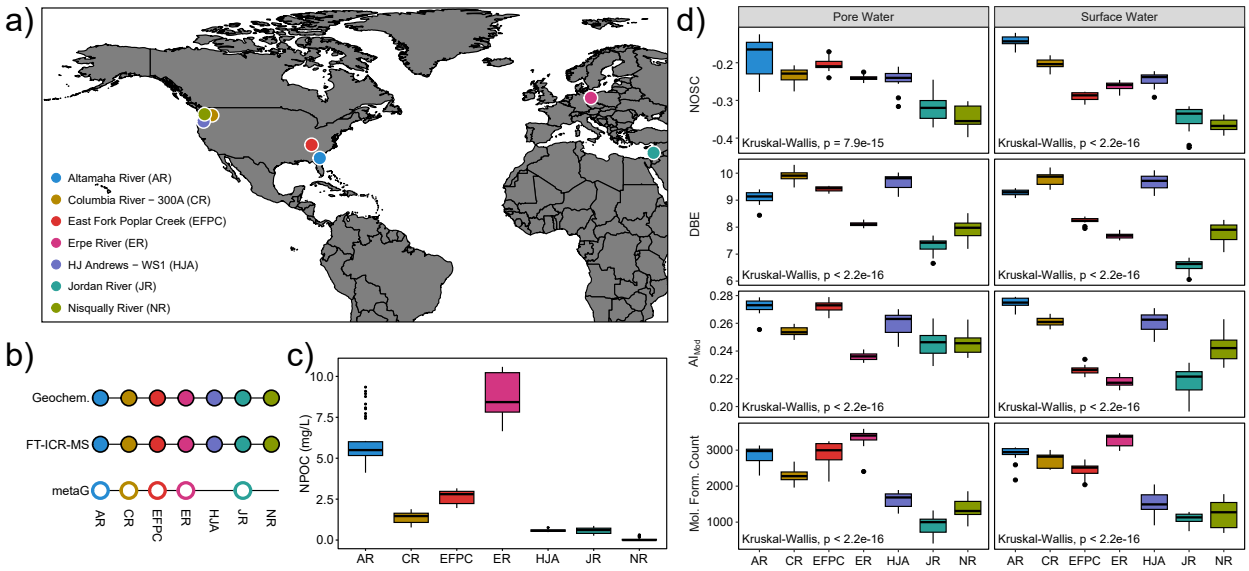



**Figure 1**: Overview of sampling locations and dissolved organic matter (DOM) characteristics. a) Map of sampling sites. b) Sampling efforts across the seven rivers. Solid circles indicate that all samples were analyzed using the corresponding method, while empty circles indicate a subset of samples was analyzed; no circle indicates no analysis. c) Boxplot of non-purgeable organic carbon (NPOC) concentrations for each river. d) Boxplot of DOM molecular characteristics.

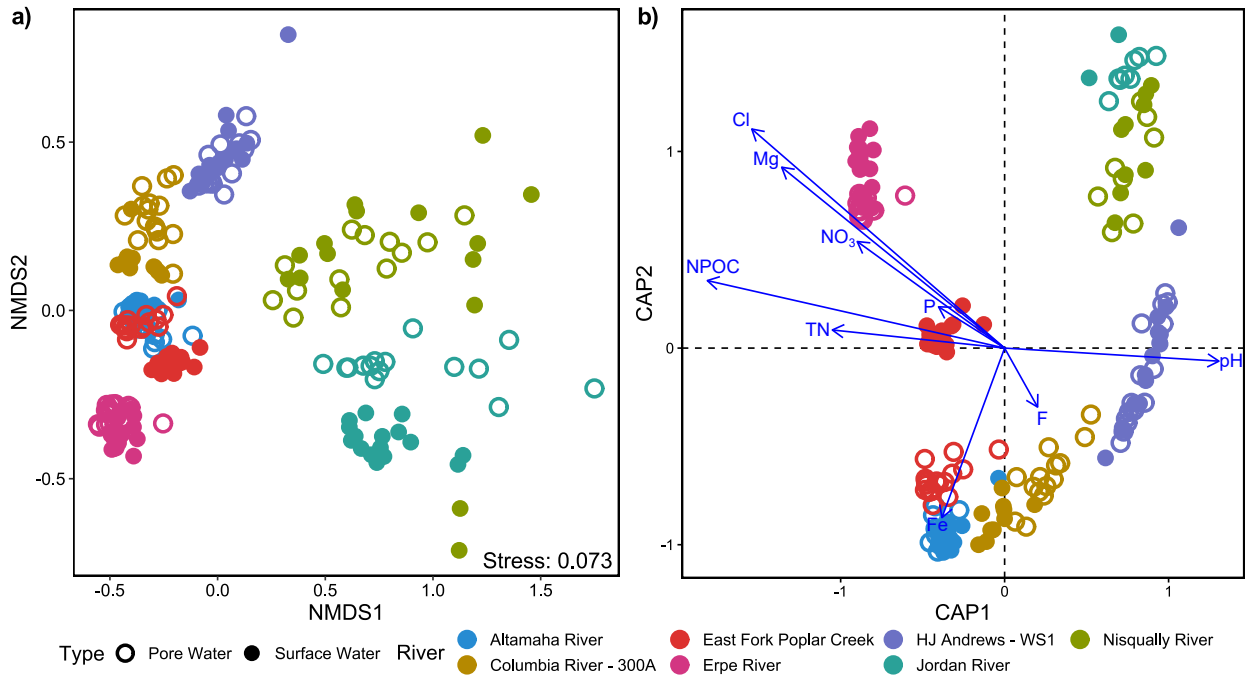



**Figure 2**: a) Non-metric multidimensional scaling (NMDS) ordination depicting the multivariate relationships of DOM across all seven rivers. b) A redundancy analysis (RDA) optimized to evaluate to what degree the geochemistry is related to DOM composition. Loading arrows indicate geochemical variables significantly related to DOM.

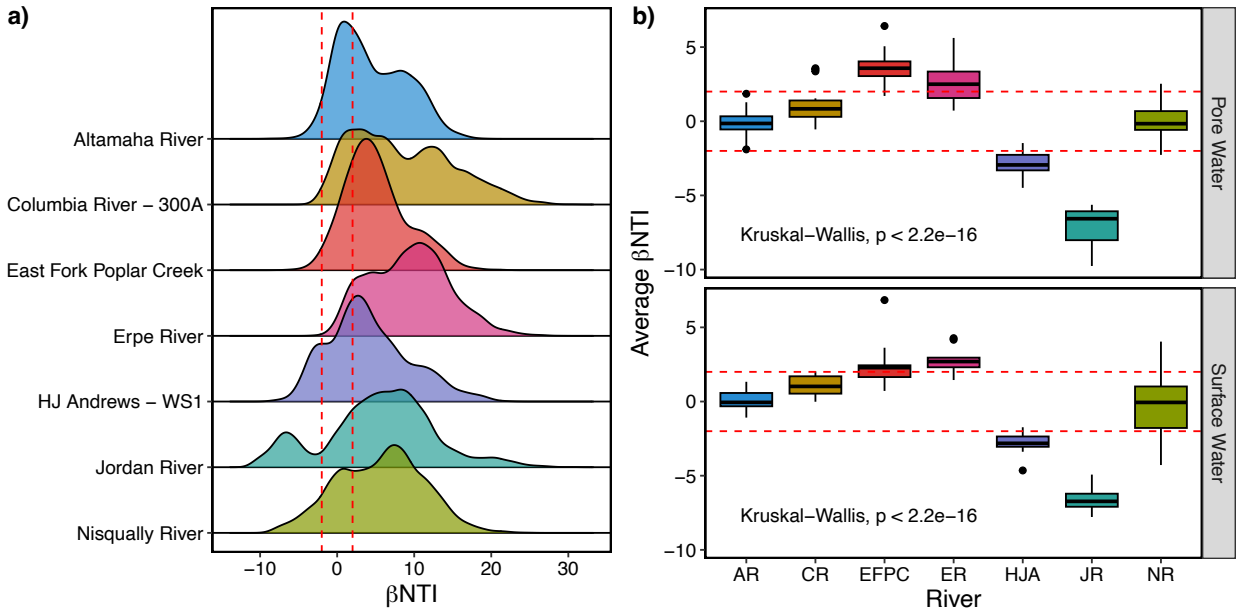



**Figure 3**: Null modeling results. a) Density plots illustrating the distribution of raw pairwise βNTI values from each River. b) Average βNTI values calculated for pairwise comparisons within the surface water (top panel) or within pore water (bottom panel).

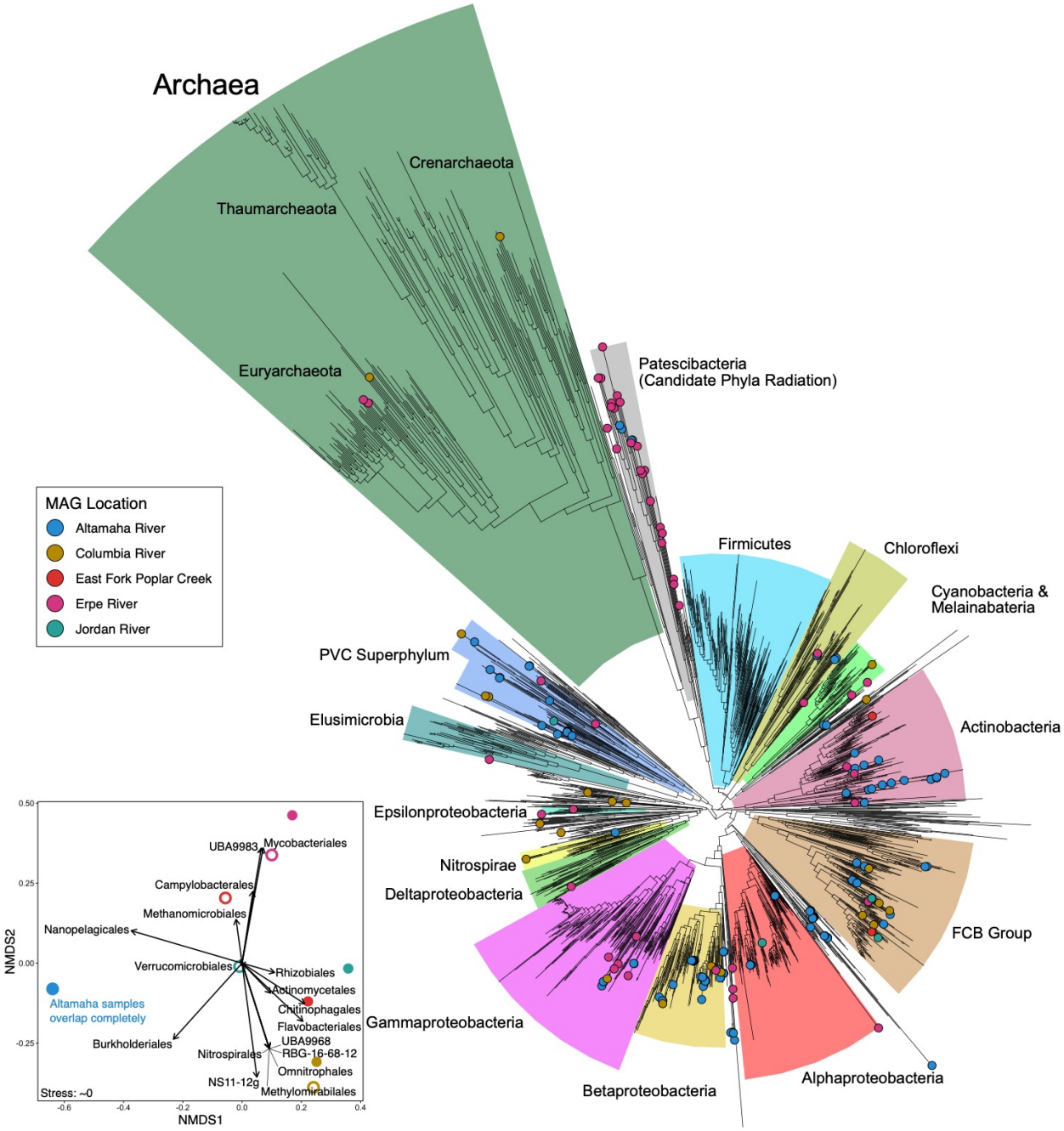



**Figure 4**: Microbial relationships. a) An rpsC phylogenetic tree generated using FastTree. Points on the tree are genomes obtained from this study and are colored based upon the river from which they were observed. b) Non-metric multidimensional scaling (NMDS) ordination indicating the beta-diversity of each sample. Loadings indicate microbial orders significantly related to the ordination structure.

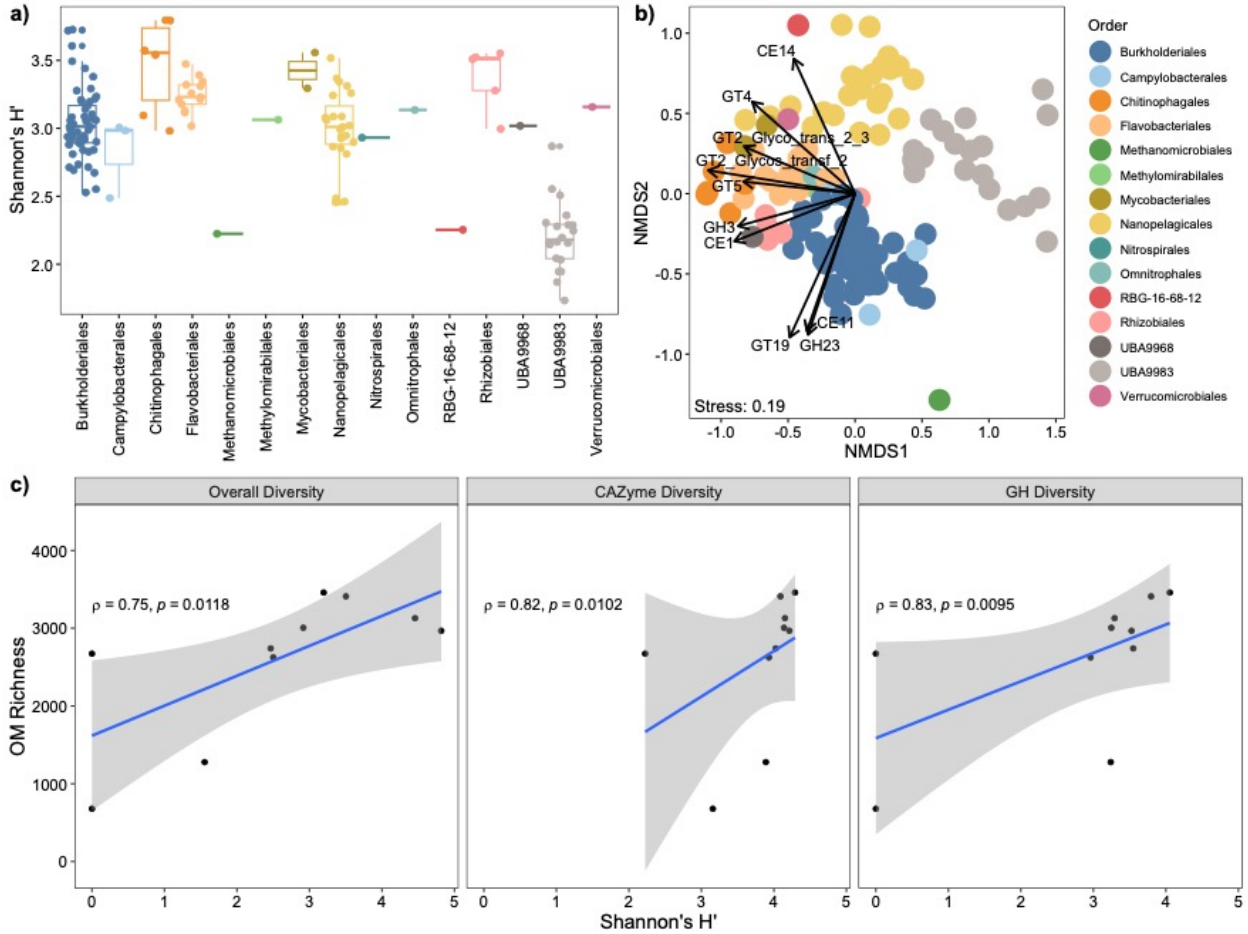
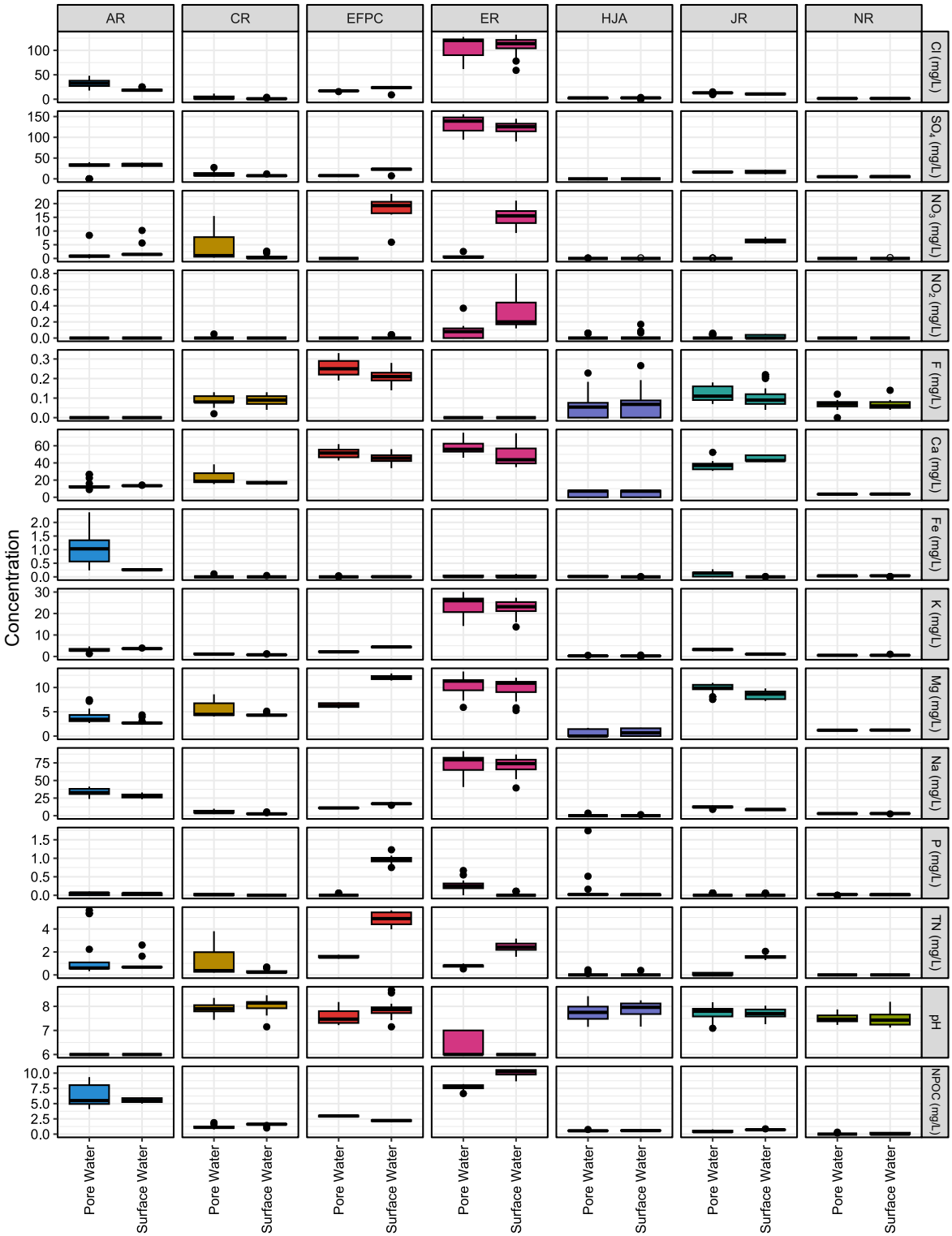



**Figure 5**: CAZyme analysis. a) Alpha-diversity analysis illustrating the CAZymes pools vary across microbial Orders. b) Non-metric multidimensional scaling (NMDS) ordination of the CAZyme pools illustrating multivariate dynamics across microbial Orders. c) Three different microbially derived diversity metrics plotted against molecular richness.

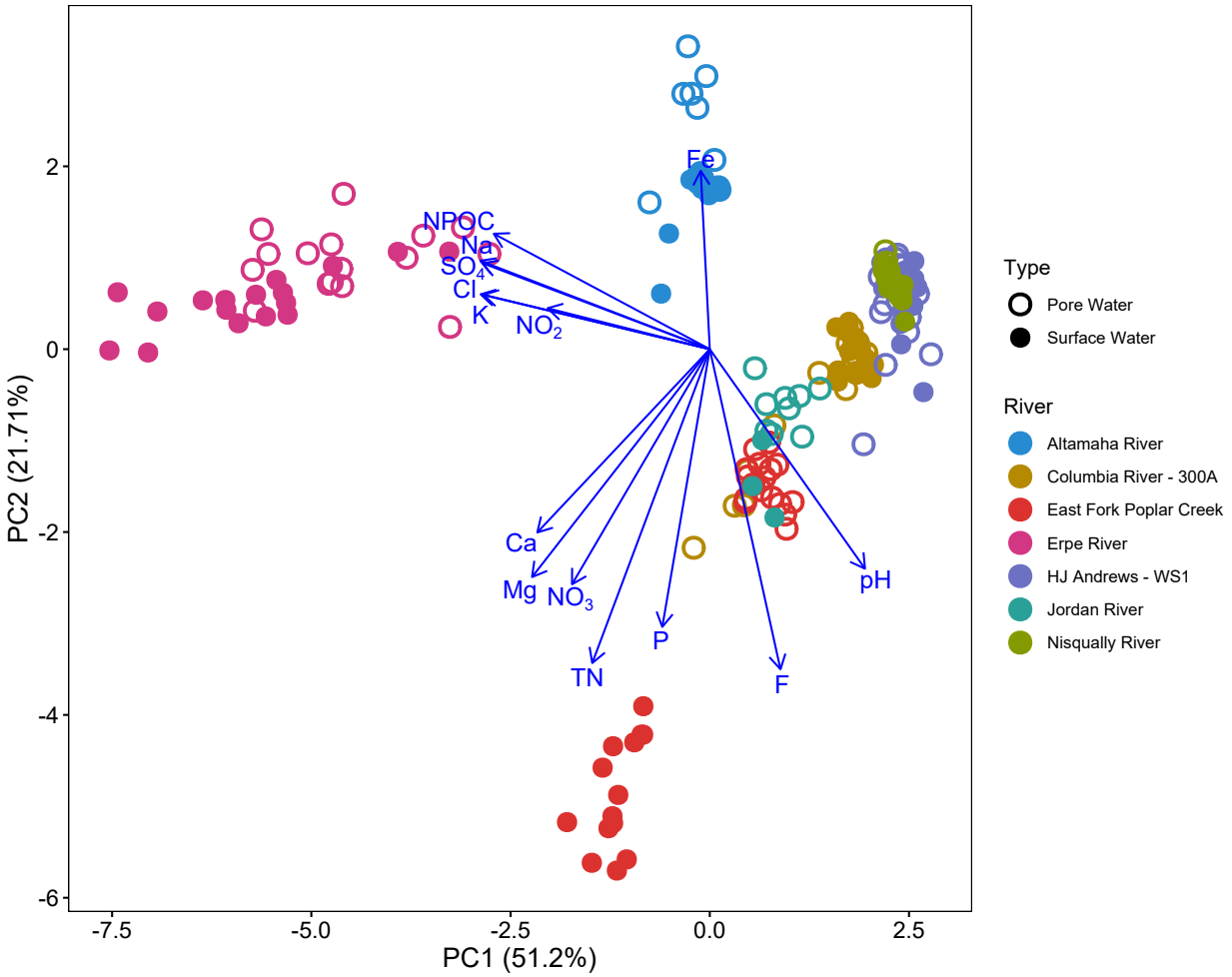
| River | Abbreviation | Location | River Type | pH | NPOC (mg/L) |
|---|---|---|---|---|---|
| Altamaha River | AR | Georgia, United States | Large River | 6 (0) | 5.98 (1.34) |
| Columbia River | CR | Washington, United States | Large River | 7.97 (0.3) | 1.38 (0.32) |
| East Fork Poplar Creek | EFPC | Tennessee, United States | Wastewater | 7.73 (0.37) | 2.60 (0.40) |
| Erpe River | ER | Berlin, Germany | Wastewater | 6.16 (0.37) | 9.5 (1.39) |
| HJ Andrews Watershed 1 | HJA | Oregon, United States | Headwater | 7.79 (0.33) | 0.57 (0.06) |
| Jordan River | JR | Israel | Headwater | 7.72 (0.26) | 0.57 (0.18) |
| Nisqually River | NR | Washington, United States | Headwater | 7.50 (0.25) | 0.06 (0.11) |

**Table 1**: Summary information for each river. Average pH and non-purgeable organic carbon are provided as contextual information (standard deviations are in parentheticals).

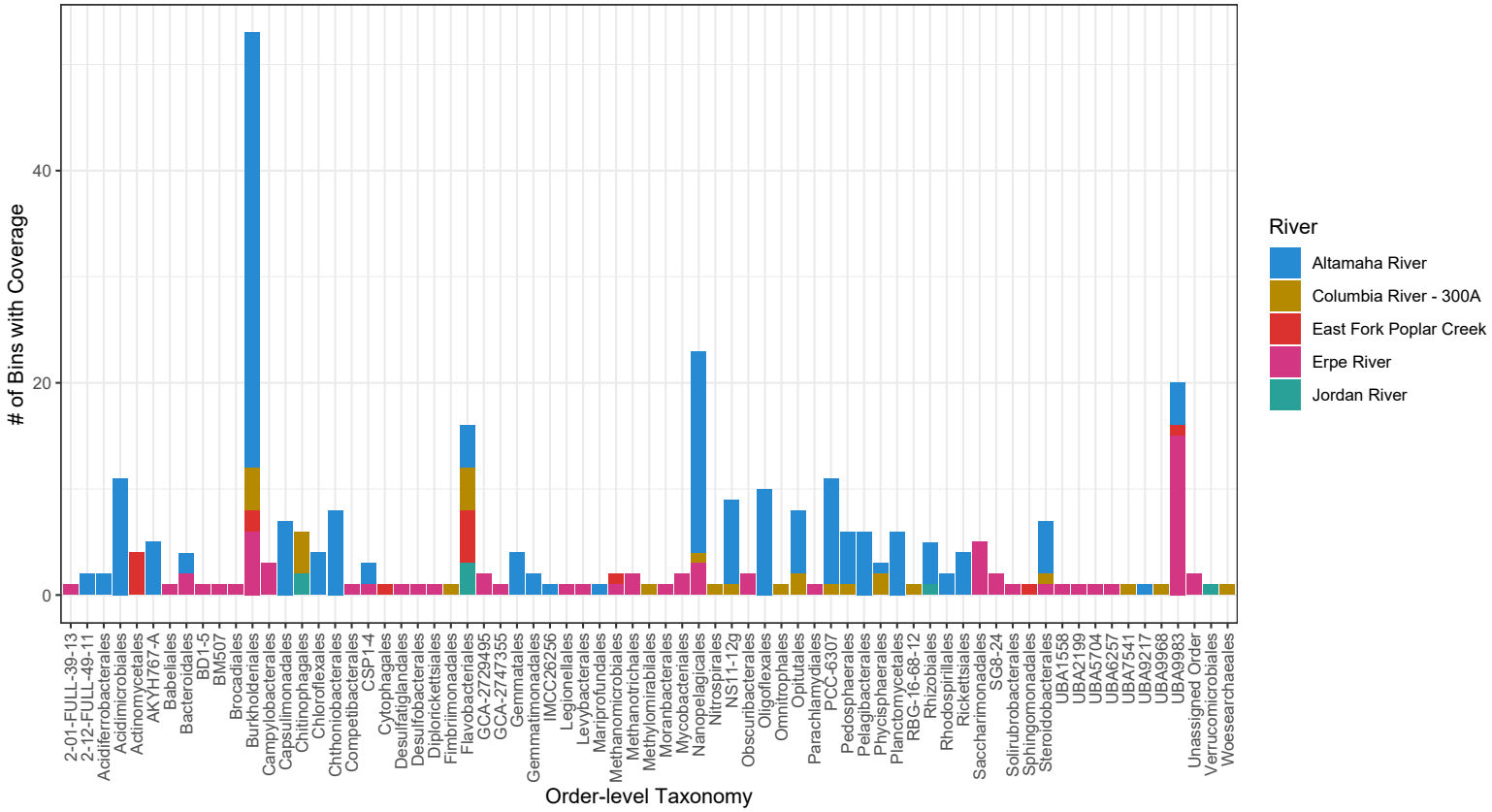
## Supplemental Information



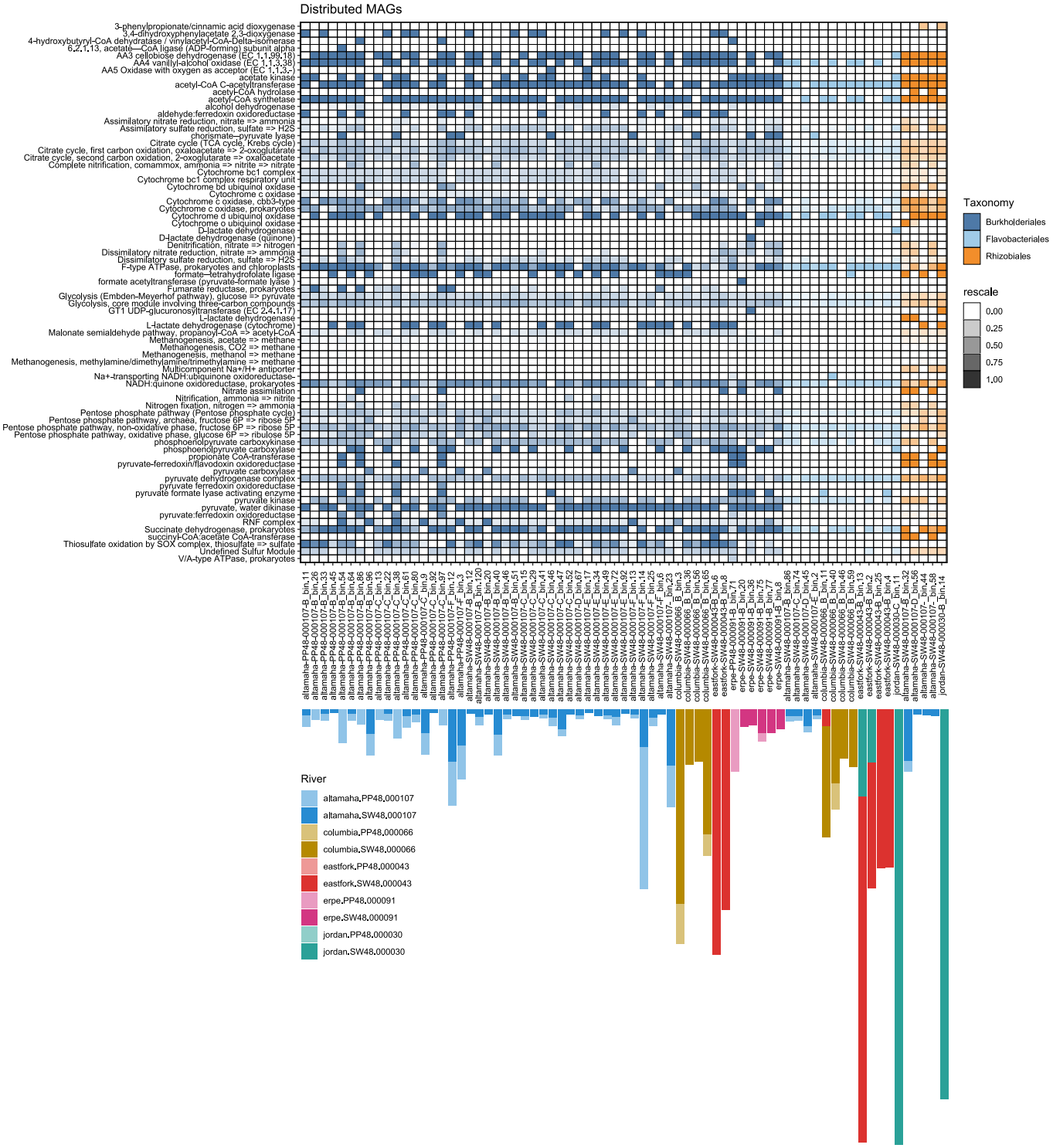

**Supplemental Figure 1:** Boxplots displaying variation in geochemical parameters in each river, separated by river type. River abbreviations are AR = Altamaha River, CR = Columbia River, EFPC = East Fork Poplar Creek, ER = Erpe River, HJA = HJ Andrews Watershed 1, JR = Jordan River, and NR = Nisqually River.



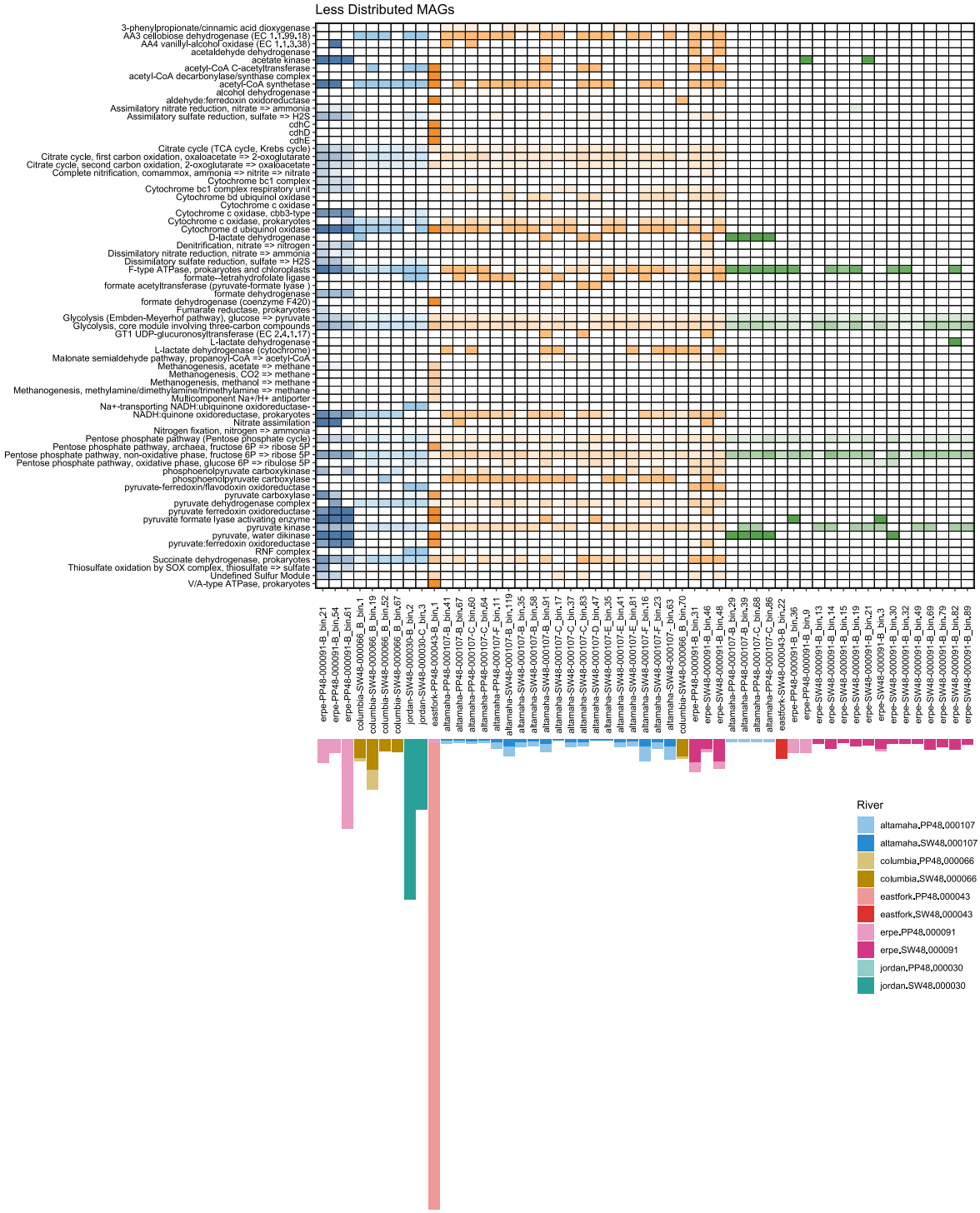

**Supplemental Figure 2:** Principal component analysis (PCA) of river and pore water geochemistry.




**Supplemental Figure 3**: Bar plot depicting the number of bins with a given GTDB taxonomy detected in each river.




**Supplemental Figure 4:** Metabolic information and abundance of broadly distributed MAGs. A heatmap showing the presence (and relative completion) of the pathway listed on the y-axis in the metagenomic assembled genome (MAG) on the x-axis. Cells in the heatmap are colored based upon the taxonomy of the MAG listed on the x-axis. Below each MAG is a visual representation of the read mapping abundance in each river with rivers indicated by the colored bars.




**Supplemental Figure 5:** Metabolic information and abundance of MAGs found in fewer rivers (the inverse of Supplemental Figure 4). A heatmap showing the presence (and relative completion) of the pathway listed on the y-axis in the metagenomic assembled genome (MAG) on the x-axis. Cells in the heatmap are colored based upon the taxonomy of the MAG listed on the x-axis. Below each MAG is a visual representation of the read mapping abundance in each river with rivers indicated by the colored bars.

**Supplemental Table 1:** Table describing the summary statistics for each metagenomic assembled genome, including (but not limited to) measurements like completion/contamination percentage, MAG size in base pairs, and N50.

**Supplemental Table 2**: Pairwise Mann Whitney U statistics for DOM measurements across rivers and sample type.

**Supplemental Table 3:** Table of significant correlation values for βNTI and geochemical relationships.

**Supplemental Table 4:** Table of pairwise Mann Whitney U statistics for average βNTI values across rivers.