Nutrient dynamics and stream order influence microbial community patterns along a 2914 km transect of the Mississippi River

Michael W. Henson¹, Jordan Hanssen², Greg Spooner², Patrick Fleming², Markus Pukonen², Frederick Stahr³, and J. Cameron Thrash¹*

¹ Department of Biological Sciences, Louisiana State University, Baton Rouge, LA 70803, U.S.A.
² O.A.R. Northwest, Seattle, WA 98103, U.S.A.
³ School of Oceanography, University of Washington, Seattle, WA 98195, U.S.A.

*Correspondence:
J. Cameron Thrash
Louisiana State University
Department of Biological Sciences
202 Life Sciences Bldg.
Baton Rouge, LA 70803
Phone: 225-578-8210
Fax: 225-578-2597
thrashc@lsu.edu

Running title: Microbial regime changes on the Mississippi River

Keywords: Mississippi River, WGCNA, microbial ecology, eutrophication
Abstract

Draining 31 states and roughly 3 million km², the Mississippi River (MSR) and its tributaries constitute an essential resource to millions of people for clean drinking water, transportation, agriculture, and industry. Since the turn of the 20th century, MSR water quality has continually rated poorly due to human activity. Acting as first responders, microorganisms can mitigate, exacerbate, and/or serve as predictors for water quality, yet we know little about their community structure or ecology at the whole river scale for large rivers. We collected both biological (16S and 18S rRNA gene amplicons) and physicochemical data from 38 MSR sites over nearly 3000 km from Minnesota to the Gulf of Mexico. Our results revealed a microbial community composed of similar taxa to other rivers but with unique trends in the relative abundance patterns among phyla, OTUs, and the core microbiome. Furthermore, we observed a separation in microbial communities that mirrored the transition from an 8th to 10th Strahler order river at the Missouri River confluence, marking a different start to the lower MSR than the historical distinction at the Ohio River confluence in Cairo, IL. Within MSR microbial assemblages we identified subgroups of OTUs from the phyla Acidobacteria, Bacteroidetes, Oomycetes, and Heterokonts that were associated with, and predictive of, the important eutrophication nutrients nitrate and phosphate. This study offers the most comprehensive view of MSR microbiota to date, provides important groundwork for higher resolution microbial studies of river perturbation, and identifies potential microbial indicators of river health related to eutrophication.
Introduction

By connecting terrestrial, lotic, and marine systems, rivers perform vital roles in both the transport and processing of compounds in all major global biogeochemical cycles (Richey et al. 2002; Ensign and Doyle 2006; Withers and Jarvie 2008; Battin et al. 2009; Savio et al. 2015).

Within the carbon cycle alone, rivers collectively discharge organic carbon to the oceans at over 0.4 Pg C yr\(^{-1}\) (Cauwet et al. 2002). Perhaps more importantly, rivers are generally net heterotrophic (Cole et al. 2007), indicating that they not only transport organic matter but host active metabolic processing of it as well. Conservative estimates place heterotrophic output of the world’s fluvial networks (streams, rivers, and estuaries) at 0.32 Pg C yr\(^{-1}\) (Cole and Caraco 2001; Battin et al. 2009). Although rivers contain a small minority of global fresh water at any given moment, the considerable volumes that pass through these systems make them relevant to models attempting to quantify global elemental transformations. However, despite the fact that microbial functions likely play a vital role in ecosystem health for both rivers themselves and their places of discharge, microbial functions in rivers remain understudied.

At 3734 km, the Mississippi River (MSR) is the fourth longest on earth, draining 31 U.S. states and two Canadian provinces- a watershed consisting of 41% of the continental U.S (Turner and Rabalais 2003; Dagg et al. 2004). The MSR is a major source of drinking water for many U.S. cities; a critical thoroughfare for transportation, commerce, industry, agriculture, and recreation; and conveys the vestiges of human activity to the Gulf of Mexico (GOM). In New Orleans, the average flow rate is over 600,000 cubic feet s\(^{-1}\) (cfs) (Rabalais et al. 1996), but can exceed 3 million cfs during flood stages (Singh 2012), and carries over 150 x 10\(^9\) kg of suspended sediment into the northern GOM annually (Dagg et al. 2004, 2005). This massive discharge includes excess nutrients (nitrogen and phosphorus), primarily from agricultural runoff.
(McIsaac et al. 2001; Turner and Rabalais 2004; Schilling et al. 2010; Duan et al. 2014; Staley et al. 2014a), that fuels one of the largest marine zones of seasonal coastal hypoxia in the world (Rabalais et al. 2002, 2007; Bianchi et al. 2010; Bristow et al. 2015). Studying microbial relationships to river eutrophication will improve our understanding of their contributions to either mitigating or exacerbating nutrient input.

Far from a homogenous jumble of organisms ferried downriver, microbial community composition changes with distance from the river mouth and/or from the influence of tributaries (Lemke et al., 2009; Winter et al., 2007; Kolmakova et al., 2014) (Read et al. 2015; Savio et al. 2015), resulting from altered nutrient concentrations (Staley et al. 2014b; Van Rossum et al. 2015; Meziti et al. 2016), differing dissolved organic matter (DOM) sources (Ruiz-González et al. 2013; Blanchet et al. 2016), and land use changes (Staley et al. 2014a; Van Rossum et al. 2015; Zeglin 2015; Meziti et al. 2016). Past studies of the Thames, Danube, Yenisei, and Columbia Rivers have found that planktonic river microbial assemblages were dominated by the phyla Actinobacteria, Proteobacteria, and Bacteriodetes, and taxa such as acl Actinobacteria, Polynucleobacter spp., GKS9 and LD28 Betaproteobacteria, CL500-29 Bacteriodetes, LD12 SAR11 Alphaproteobacteria, and Novosphingobium spp. (Zwart et al. 2002; Savio et al. 2015).

More recent 16S rRNA gene amplicon and metagenomic studies of the Minnesota portion of the MSR corroborated previous research in other rivers that identified an increased proportion (Savio et al. 2015) or richness (Read et al. 2015) of freshwater taxa with river distance, and an increased abundance of “core” river microbes (Staley et al. 2013) with cumulative residence time (Savio et al., 2015, Winter et al., 2007; Read et al., 2015; Staley et al., 2013; Crump et al., 1996).

Researchers have suggested that these patterns supported the application of the River Continuum Concept (RCC) (Vannote et al., 1980) to river microbiota. The RCC postulates that
as a river increases in size, the influences of riparian and other inputs will decrease as the river
establishes a dominant core community. Richness will increase with stream order complexity
before decreasing in higher order rivers (Vannote et al., 1980). Therefore, as continuous systems
with increasing volumes and residence times, river microbiota should transition from
experiencing strong influences of mass effects from terrestrial and tributary sources to systems
where species sorting plays a more important role (Fortunato et al. 2012; Besemer et al. 2013;
Savio et al. 2015). Analysis of MSR microbial communities suggests the importance of point
source inputs on local communities while regional scale differences are more influenced by mass
immigration from sediment and tributaries (Staley et al. 2013, 2016; Payne et al. 2017).

Complicating matters, particle-associated communities in rivers (frequently defined as
those found on filters of > ~3 μm) remain distinct from their free-living counterparts (Jackson et
al., 2014; Crump et al., 1999; Riemann and Winding, 2001), potentially due to increased
production rates from access to readily obtainable carbon (Crump et al., 1999; Crump and
Baross, 1996). Typical taxa associated with particles include OTUs related to the Bacteroidetes
clades Flavobacteria and Cytophaga, Planctomycetes, Rhizobium spp., and Methylphilaceae
tspp. (Crump and Baross 1996; Allgaier and Grossart 2006; D’Ambrosio et al. 2014; Jackson et
al. 2014). However, consistent trends in particle community composition are murky, with recent
evidence suggesting organisms may switch between free-living and particle-associated lifestyles
depending on substrate availability and chemical queues (Grossart 2010; D’Ambrosio et al.
2014). Thus, rivers constitute complex and highly dynamic ecosystems from a metacommunity
perspective.

Although important insights have been gained from recent research on portions of the
MSR (Staley et al. 2013, 2015, 2016; Payne et al. 2017), microbiological transects at the whole-
river or catchment scale have yet to be completed. This study aimed to i) compare size-
fractionated MSR microbial community populations to those in other rivers, ii) examine within-
river heterogeneity of microbial communities, and iii) identify MSR microorganisms most
strongly associated with eutrophication— all at a near whole-river scale. During the fall of 2014,
we completed the most extensive microbiological survey of the Mississippi River to date with a
continual rowed transect over two months (70 days). Rowers from the adventure education non-
profit OAR Northwest collected samples from Minneapolis, MN to the Gulf of Mexico (2918
km) (Fig. 1A). Our findings expand the current information available on microbial assemblages
in major lotic ecosystems; further delineate the relationships between microbial structure and
stream order, nutrients, and volume; and identify MSR taxa predictive of the eutrophication
nutrients nitrate and phosphate.

Materials and Methods

Sampling and Cell Counts

We used rowboats and a simple filtration protocol (Supplementary Information) to collect water
from 39 sites along a continually rowed transect of the MSR, starting in Lake Itasca and ending
in the GOM, over 70 days from September 18th to November 26th, 2014. Sites were chosen to be
near major cities and above and below large tributaries. After some samples were removed due
to insufficient sequence data, contamination, or incomplete metadata (see below), the final
usable set of samples included 38 sites starting at Minneapolis (Fig. 1A, Table S1). Most
sampling occurred within the body of the river, although due to safety issues, three samples were
collected from shore (Table S1). We collected duplicate samples at each site, but because
separate rowboat teams frequently collected these sometimes several dozen meters apart, they cannot be considered true biological replicates and we have treated them as independent samples. At each site, we filtered 120 mL of water sequentially through a 2.7 µm GF/D filter (Whatman GE, New Jersey, USA) housed in a 25 mm polycarbonate holder (TISCH, Ohio, USA) followed by a 0.2 µm Sterivex filter (EMD Millipore, Darmstadt, Germany) with a sterile 60 mL syringe (BD, New Jersey, USA). We refer to fractions collected on the 2.7 µm and 0.22 µm filters as > 2.7 µm and 0.2-2.7 µm, respectively. Flow-through water from the first 60 mL was collected in autoclaved acid-washed 60 mL polycarbonate bottles. Both filters were wrapped in parafilm, and together with the filtrate, placed on ice in Yeti Roadie 20 coolers (Yeti, Austin, TX) until shipment to LSU. Further, 9 mL of whole water for cell counts was added to sterile 15 mL Falcon tubes containing 1 mL of formaldehyde and placed into the cooler. We monitored cooler temperature with HOBO loggers (Onset, Bourne, MA) to ensure samples stayed at ≤ 4ºC. The final cooler containing samples from sites P-Al had substantial ice-melt. Though our filters were wrapped in parafilm, we processed melted cooler water alongside our other samples to control for potential contamination in these filters. Given that some of our samples were expected to contain low biomass, we also included duplicate process controls for kit contamination (Salter et al. 2014; Weiss et al. 2014) with unused sterile filters. Flow-through 0.2 µm filtered water from each collection was analyzed for SiO$_4$$^-$, PO$_4$$^{3-}$, NH$_4$$^+$, NO$_3$$^-$, and NO$_2$$^-$ (µg/L) at the University of Washington Marine Chemistry Laboratory (http://www.ocean.washington.edu/story/Marine+Chemistry+Laboratory). Aboard-rowboat measurements were taken for temperature and turbidity. We determined turbidity by deploying a secchi disk (Wildco, Yulee, FL), while drifting with the current so the line hung vertically. It was lowered until no longer visible, then raised until just visible, and measured for its distance below
the waterline. We then calculated secchi depth from the average of two measurements.

Temperature was measured with probes from US Water Systems (Indianapolis, IN), rinsed with distilled water between samples. Samples for cell counts were filtered through a 2.7 µm GF/D filter, stained with 1x Sybr Green (Lonza), and enumerated using the Guava EasyCyte (Millipore) flow cytometer as previously described (Thrash et al. 2015).

**DNA extraction and Sequencing**

DNA was extracted from both filter fractions and controls using a MoBio PowerWater DNA kit (MoBio Laboratories, Carlsbad, CA) following the manufacturer’s protocol with one minor modification: in a biosafety cabinet (The Baker Company, Stanford, ME), Sterivex filter housings were cracked open using sterilized pliers and filters were then removed by cutting along the edge of the plastic holder with a sterile razor blade before being placed into bead-beating tubes. DNA was eluted with sterile MilliQ water, quantified using the Qubit2.0 Fluorometer (Life Technologies, Carlsbad, CA), and stored at -20° C. Bacterial and archaeal sequences were amplified at the V4 region of the 16S rRNA gene using the 515f and 806r primer set (Caporaso et al. 2012), and eukaryotic sequences from the V9 region of the 18S rRNA gene using the 1391r and EukBR primer set (Amaral-Zettler et al. 2009). Amplicons were sequenced on an Illumina MiSeq as paired-end 250 bp reads at Argonne National Laboratory. Sequencing of the 16S and 18S rRNA gene amplicons resulted in 13,253,140 and 13,240,531 sequences, respectively.

**Sequence Analysis**
We analyzed amplicon data with Mothur v.1.33.3 (Schloss et al. 2009) using the Silva v.119 database (Pruesse et al. 2007; Quast et al. 2013). Briefly, 16S and 18S rRNA gene sequences were assembled into contigs and discarded if the contig had any ambiguous base pairs, possessed repeats greater than 8 bp, or were greater than 253 or 184 bp in length, respectively. Contigs were aligned using the Silva rRNA v.119 database, checked for chimeras using UCHIME (Edgar et al. 2011), and classified also using the Silva rRNA v.119 database. Contigs classified as chloroplast, eukaryotes, mitochondria, or “unknown;” or as chloroplast, bacteria, archaea, mitochondria, or “unknown;” were removed from 16S or 18S rRNA gene data, respectively. The remaining contigs were clustered using the cluster.split() command into operational taxonomic units (OTUs) using a 0.03 dissimilarity threshold (OTU0.03). After these steps, 146,725 and 131,352 OTUs remained for the 16S and 18S rRNA gene communities, respectively.

Sample quality control

To evaluate the potential for contamination from extraction kits, cooler water in the last set of samples, or leaking/bursting of pre-filters, all samples were evaluated with hierarchical clustering and NMDS analysis. Hierarchical clustering was preformed in R using the hclust function with methods set to “average”, from the vegan package (Oksanen et al. 2015). Samples were removed from our analysis if they were observed to be outliers in both the NMDS and hierarchical clustering such that they grouped with our process controls. The process and cooler water controls were extreme outliers in both, as was sample L2 (Fig. S1, S2). Sterivex and prefilter samples generally showed strong separation with the exception of three 16S rRNA gene samples- STER X2, W2, S2 (Fig. S1, S2). The only other samples that were removed were due to missing chemical data (Lake Itasca1-2, A1-2) or failed sequencing (16S STER Af1; 16S PRE
S2, X2; 18S PRE O1). Not including process or cooler water controls, 152 samples were sequenced each for prokaryotic and eukaryotic communities. After these QC measures, 144 and 149 samples remained in the analyses from the 16S and 18S rRNA gene amplicons, respectively. Further, to control for potential contaminants, any OTU with greater 20 reads in the process or cooler controls was removed from the data set. 146,725 and 131,327 OTUs remained after these steps for 16S and 18S rRNA gene communities, respectively.

**Alpha and Beta Diversity**

OTU$_{0.03}$ analyses were completed with the R statistical environment v.3.2.1 (R Development Core Team 2015). Using the package PhyloSeq (McMurdie and Holmes 2013), alpha-diversity was first calculated on the unfiltered OTUs using the “estimate richness” command within PhyloSeq, which calculates Chao1 (McMurdie and Holmes 2013). After estimating chao1, potentially erroneous rare OTUs, defined here as those without at least two sequences in 20% of the data, were discarded. After this filter, the dataset contained 950 and 724 16S and 18S rRNA gene OTUs, respectively. For site-specific community comparisons, OTU counts were normalized using the package DESeq2 (Love et al. 2014) with a variance stabilizing transformation (Learman et al. 2016). Beta-diversity between samples was examined using Bray-Curtis distances via ordination with non-metric multidimensional scaling (NMDS). Analysis of similarity (ANOSIM) was used to test for significant differences between groups of samples (e.g lower versus upper MSR) using the anosim function in the vegan package (Oksanen et al. 2015). The influence of environmental parameters on beta-diversity was calculated in R with the envfit function.
To identify specific OTUs with strong relationships to environmental parameters (e.g. turbidity, \( \text{NO}_3^- \)), we employed weighted gene co-expression network analysis (WGCNA) (Langfelder and Horvath 2008) as previously described for OTU relative abundances (Guidi et al. 2016). First, a similarity matrix of nodes (OTUs) was created based on pairwise Pearson correlations across samples. This was transformed into an adjacency matrix by raising the similarity matrix to a soft threshold power (\( p; p = 6 \) for 16S and 18S > 2.7 \( \mu \)m, \( p = 4 \) for 16S 0.2-2.7 \( \mu \)m, \( p = 9 \) for 18S 0.2-2.7 \( \mu \)m) that ensured scale-free topology. Submodules of highly co-correlating OTUs were defined with a topological overlap matrix and hierarchical clustering. Each submodule, represented by an eigenvalue, was pairwise Pearson correlated to individual environmental parameters (Figs. S7-10A). To explore the relationship of submodule structure to these parameters, OTUs within the submodule were plotted using their individual correlation to the parameter of interest (here nitrate or phosphate) and their submodule membership, defined as the number of connections within the module (Figs. S7-10B, D). Strong correlations between submodule structure and an environmental parameter facilitate identification of OTUs that are highly correlated to that parameter. To evaluate the predictive relationship between a submodule and a parameter, we employed partial least square regression (PLS) analysis. PLS maximizes the covariance between two parameters (e.g., OTU abundance and nitrate concentration) to define the degree to which the measured value (OTU abundance) can predict the response variable (nutrient concentration). The PLS model was permuted 1000 times and Pearson correlations were calculated between the response variable and leave-one-out cross-validation (LOOCV) predicted values. Modeled values were then compared with measured values to determine the explanatory power of the relationships (Figs. S7-10C, E). Relative contributions of individual
OTUs to the PLS regression were calculated using value of importance in the projection (VIP) (Chong and Jun 2005) determination. PLS was run using the R package pls (Mevik and Wehrens 2007), while VIP was run using additional code found here: http://mevik.net/work/software/VIP.R.

Environmental Ontology
Environmental ontology of individual 16S rRNA gene OTUs was determined using the SEQenv (v1.2.4) pipeline (https://github.com/xapple/seqenv; Sinclair et al. 2016) as previously described (Savio et al. 2015). Briefly, representative sequences of our OTUs were searched against the NCBI nt database (updated on 07/01/2016) using BLAST and filtered for hits with a minimum of 99% identity. From each hit, a text query of the metadata was performed to identify terms representing the sequence’s environmental source. The text was mined for EnvO terms (http://environmentontology.org/) and the frequency in which the terms appeared for each OTU was recorded. Using the seq_to_names output provided by SEQenv, EnvO terms were formed into six groups: Freshwater, Salt Water, Anthropogenic, Terrestrial, Sediment, and Unclassified (Table S1). To be assigned a group, an OTU had to have the majority (> 50%) of its hits classified to that term, while OTUs with equal distribution between two or more groups were termed as Unclassified. OTUs that returned no significant hits to an EnvO term were assigned to a seventh category, NA. OTUs and their corresponding relative abundances were merged based on the assigned group and plotted with R.

Accession numbers
Community 16S and 18S rRNA gene sequence fastq files are available at the NCBI Sequence Read Archive under the accession numbers: SRR3485674 - SRR3485971 and SRR3488881 - SRR3489315, respectively.

**Code Availability**

All code used for Mothur, SeqENV, PhyloSeq, WGCNA, and PLS regression analyses can be found on the Thrash lab website (http://thethrashlab.com/publications) with the reference to this manuscript linked to “Supplemental Information”.

**Results**

Using rowboats and a simple syringe-based filtration protocol, we measured 12 biological, chemical, and physical parameters (e.g. 16S and 18S rRNA gene communities, NH$_4^+$, river speed, etc.) from 38 sites along a 2918 km transect of the MSR (Fig. 1A). River order increases dramatically at the Missouri confluence (eighth to tenth Strahler order (Pierson et al. 2008), which corresponded to overall discharge (Fig. 1A) and beta diversity changes discussed below. Therefore, we refer to this juncture as the separator between the upper (0 km – 1042 km, Sites A-S) and lower MSR (1075-2914 km, Sites T-Al). Within the upper MSR, NO$_3^-$, PO$_4^{3-}$, and NO$_2^-$ were variable but generally increased downriver until peak concentrations near the confluences of the Illinois and Missouri Rivers. This gave way to lower and more consistent concentrations along the lower MSR (Fig. 1B). Turbidity (inversely related to secchi disk visibility) increased steadily downriver to a maximum near the Illinois and Missouri River confluences (1042 km, Site S) (Fig. 1B), then trended downwards for the rest of the transect. Planktonic (< 2.7 µm) cell counts varied between 1 and 3x10$^6$ cells·mL$^{-1}$ in the upper MSR, and decreased to high 10$^5$ cells
in the lower MSR (Fig. 1B). Water temperature ranged from 19°C (133km, Site E) to 11.7°C (2552 km, Site Ag), and river speed, excluding three sites sampled from shore, was between 5.5 mph at Site Y and 0.4 mph (597 km, Site L) (Table S1). Spearman rank correlations of the measured environmental parameters showed strong positive correlations between nitrate, phosphate, distance, and increased turbidity; while nitrate and phosphate both strongly correlated negatively to water temperature and river speed (Table S1).

**Bacterial and archaeal communities**

We observed a clear distinction between the 0.2-2.7 µm and > 2.7 µm 16S rRNA gene communities (ANOSIM R = 0.65, P = 0.001) (Fig. S1A). Both size fractions had comparable species richness and evenness that trended upwards downriver (Figs. S3A, B), although a peak occurred for both fractions at sites O-Q (761-999 km) below the Des Moines River and above the Illinois River (Fig. S3A, B). Both size fractions (NMDS stress = 0.14 for each) also showed a significant separation between sites above and below the Missouri River confluence (ANOSIM, > 2.7 µm: R = 0.44, P=0.001; 0.2-2.7 µm: R = 0.48, P = 0.001) (Fig. 2A, B), that the Eukaryotic fractions mirrored (below). Phosphate and turbidity had the highest correlation with the separation between the upper and lower > 2.7 µm communities (r = 0.57, r= 0.54, respectively), with water temperature and distance (r > 0.40) also contributing (Fig. 2A). At an OTU level, taxa related to the acI clade (Actinobacteria) and unclassified *Bacillaceae* (r > 0.77, P = 0.001) contributed most to the separation between the upper and lower > 2.7 communities, with OTUs related to the *Bacillales, Gemmatimonadaceae, Peptococcaceae*, and *Micromonosporaceae* clades also a factor (r > 0.70, P = 0.001) (Figure 2A, Table S1). For the 0.2-2.7 µm fraction, distance and nitrate were the strongest correlating environmental factors with the distinction.
between upper and lower communities (r = 0.59 and r = 0.47, respectively), although phosphate, turbidity, and water temperature (r > 0.40 for each) also contributed (Fig. 2B). OTUs related to *Flavobacterium* and an unclassified Bacterium (closest NCBI BLAST hit *Acidovorax* sp., KM047473), most strongly contributed to the separation between the 0.2-2.7 µm communities (r > 0.52, P = 0.001). Other important OTUs driving the separation between upper and lower communities belonged to the clades Bacteroidetes, *Microbacteriaceae*, *Clostridiales*, and *Holophagaceae* (r > 0.49, P = 0.001) (Figure 2B, Table S1).

At the phylum level, Proteobacteria, Actinobacteria, and Bacteroidetes dominated bacterial communities in both fractions (Figs. 3A and B) (Table S1). Proteobacteria in the > 2.7 µm fraction fluctuated widely in abundance (Fig. 3A), whereas their 0.2-2.7 µm counterparts generally increased in relative abundance downriver (Fig. 3B). 0.2-2.7 µm Bacteroidetes and Actinobacteria generally decreased in the upper river and stabilized in the lower river. These phyla showed considerable abundance variation in the > 2.7 µm fraction. Cyanobacteria in the > 2.7 µm fraction negatively correlated with increased turbidity (Spearman rank = 0.67), consistent with lower irradiance. Both > 2.7 µm and 0.2-2.7 µm Acidobacteria increased in abundance downriver and positively correlated with river distance (Wilcoxon single ranked test, P = < 0.01) (Fig. 3A, B). Within the 0.2-2.7 µm fraction, the five most abundant OTUs were classified as a LD12 (OTU11), two acI clade OTUs (OTU4, OTU7), a *Limnohabitans* sp. (OTU2), and a LD28 (OTU8) (Table S1). Comparatively, an unclassified *Methylophilaceae* (OTU1), a *Planktothrix* sp. (OTU12), a NS11-12 marine group (OTU21), an *Aquirestis* sp. (OTU17), and an unclassified *Sphingobacteriales* (OTU25) were the most abundant OTUs in the > 2.7 µm fraction (Table S1). Archaea occurred at much lower relative abundances: we found only eight OTUs belonging to the phyla Euryarchaeota and Thaumarchaeota, collectively. Thaumarchaeota increased in

---

**Note:** The text above contains a preprint with the citation format: bioRxiv preprint first posted online Dec. 4, 2016; doi: http://dx.doi.org/10.1101/091512. The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.
abundance along the transect more in the > 2.7 μm fraction compared to the 0.2-2.7 μm (Fig. 3A, B). In both fractions, we only detected Euryarchaeota at very low abundances. Importantly, the primers used in this study may miss some archaeal taxa (Parada et al. 2015).

Microbial eukaryotic communities

Eukaryotic communities, observed via the 18S rRNA gene, also showed a significant separation between > 2.7 μm and 0.2-2.7 μm fractions (ANOSIM R = 0.689, P = 0.001) (Fig. S1B). As expected due to generally larger cell sizes in microbial eukaryotes compared to prokaryotes, species richness remained higher in the > 2.7 μm vs. 0.2-2.7 μm fractions (Fig. S3C, D). Richness in the > 2.7 μm fraction gradually increased downriver, similarly to prokaryotic communities, but remained relatively stable within the 0.2-2.7 μm fraction. Both the > 2.7 μm (stress = 0.113) and 0.2-2.7 μm (stress = 0.146) fractions also showed a significant separation between the upper and lower MSR (ANOSIM, > 2.7 μm: R = 0.696, P = 0.001; 0.2-2.7 μm: R = 0.576, P = 0.001) (Fig. 2C, D). Distance and phosphate were the top two environmental factors influencing this distinction (r = 0.75, r = 0.48; r = 0.70, r = 0.54, respectively) (Fig. 2C, D; Table S1). No other factors had correlations > 0.40 (Table S1). At the OTU level, taxa related to an unclassified Ochrophyta (OTU63) and an unclassified Eukaryote (OTU1) separated the 0.2-2.7 μm communities (r > 0.63, P= 0.001), while the same unclassified Eukaryote OTU (OTU1) and a second unclassified Eukaryote (OTU222) contributed most to separating the > 2.7 μm communities (r > 0.80, P = 0.001) (Figure 2C and D, Table S1).

Stramenopiles (or Heterokonts), encompassing diatoms and many other forms of algae, and OTUs that could not be classified at the phylum level dominated both the > 2.7 μm and 0.2-2.7 μm communities (Fig. 4). Stramenopiles accounted for over 25% of both communities, with
higher abundances in the upper vs. lower river. We observed a similar trend of disparate
abundances between the upper and lower river for > 2.7 μm Cryptomonadales and 0.2-2.7 μm
Nucletmycea, the latter of which include fungi (Fig. 4A; Table S1). Within the 0.2-2.7 μm
fraction, we identified the five most abundant OTUs as three unclassified Bacillariophytina
(OTU7, OTU14, OTU9), a Pythium sp. (OTU170), and an unclassified Cryptomonas (OTU11)
(Table S1). Comparatively, two unclassified Eukaryotes (OTU2 and OTU1), an unclassified
Stramenopiles (OTU3), an unclassified Perkinsidae (OTU13), and an unclassified
Chrysophyceae (OTU6) had the highest abundance in the > 2.7 μm fraction (Table S1).

The Mississippi River Core Microbiome
We defined the core microbiome as those OTUs detectable after normalization in greater than
90% of the samples. The 16S rRNA gene > 2.7 μm and 0.2-2.7 μm core microbiomes consisted
of 82 and 98 OTUs, respectively, classified into eight different phyla: Proteobacteria,
Actinobacteria, Bacteroidetes, Cyanobacteria, Verrucomicrobia, Chloroflexi, Chlorobi,
Gemmatimonadetes- and composed of taxa such as freshwater SAR11 (LD12), Limnohabitans
sp., Polynucleobacter sp., acl clade, LD28 clade, and Planktothrix sp. (Table S1). Core
microbome relative abundance in both fractions decreased along the upper river but stabilized in
the lower river (Fig. 5A). We confirmed this effect by analyzing the upper and lower core
microbiomes separately. Although the total OTU numbers changed (81 and 116 OTUs in the
upper MSR and 160 and 144 OTUs in the lower MSR for the > 2.7 μm and 0.2-2.7 μm fractions,
respectively), the trends remained the same (Fig. S4).

Eighty OTUs comprised the > 2.7 μm 18S rRNA gene core microbiome (Fig. 5B). We
classified these as Alveolata, Cryptophceae, Nucletmycea, Stramenopiles, or unclassified
Eukaryota (Table S1). Again, consistent with larger organism sizes, and thus fewer OTUs overall, the 0.2-2.7 μm Eukaryotic core microbiome comprised only 21 OTUs (Fig. 5B). These OTUs consisted of Alveolata, Nucleomycota, Stramenopiles, or unclassified Eukaryota (Table S1). While the 0.2-2.7 μm core microbiome remained relatively stable along the river, the >2.7 μm core decreased along the upper MSR before stabilizing in the lower river, similarly to that of the prokaryotes (Fig. 5B).

16S rRNA gene environmental ontology

We successfully classified 136 of the 950 16S rRNA gene OTUs with EnvO terminology (Table S1). Freshwater organisms dominated the OTU designations that we could assign, although their relative abundance in both the >2.7 μm and 0.2-2.7 μm fractions decreased along the upper river before stabilizing in the lower MSR (Fig. S5A, B). However, LD12 (OTU11), the most abundant OTU in our dataset and a well-established freshwater organism (Zwart et al. 2002; Newton et al. 2011; Salcher et al. 2011), did not receive an EnvO classification, demonstrating the limitations of this technique with current database annotations. Terrestrial organisms from the >2.7 μm fraction decreased along the transect (Fig. S5A), while those in the 0.2-2.7 μm fraction remained stable (Fig. S5B). Although representing a minor fraction of total OTUs, sediment-associated >2.7 μm microorganisms fluctuated widely in abundance along the upper river before stabilizing in the lower, whereas those in the 0.2-2.7 μm fraction occurred in similar abundances throughout the river. Taxa associated with anthropogenic sources in both fractions increased along the river (Fig. S5A-B).

Network analyses identify taxa associated with and predictive of eutrophication
We applied Weighted Gene Correlation Network Analysis (WGCNA) to identify co-occurring groups of OTUs (submodules) that also had significant associations with the eutrophication nutrients phosphate and nitrate. Of the submodules identified through WGCNA as being most strongly correlated to phosphate and nitrate, we restrict our discussion to those modeled via PLS analysis to predict > 50% of the measured nutrient concentration. Our additional PLS analyses for those submodules predicting < 50% of nitrate and/or phosphate concentrations are included in figure S6.

Three submodules in the prokaryotic and eukaryotic fractions were strongly associated with phosphate. The 0.2-2.7 µm prokaryotic submodule most associated with phosphate was composed of 51 OTUs (Table 1), and had moderate correlation between the submodule structure and phosphate (Fig. S7D). Strong correlations between submodule structure and the measured nutrient suggests that individual submodule OTUs that also have strong correlations to the nutrient are the most important organisms associated with that nutrient (Langfelder and Horvath 2008). PLS modeling determined that this prokaryotic submodule predicted 80% of measured phosphate concentrations (Table 1). Variable importance in the projection (VIP) analysis found OTUs corresponding to an unclassified *Holophagaceae*, an unclassified *Gemmatimonadaceae*, and an unclassified *Burkholderiaceae* were the three most important in the PLS model for phosphate (Fig. 6A, Table S1). OTU322 (Acidobacteria subgroup 6), had moderate correlation to phosphate but high node centrality (n = 51; Fig. 6A, Table S1), corroborating evidence that freshwater sediment Acidobacteria subgroup 6 often occur in co-culture with Alphaproteobacteria and may be metabolically connected (Kielak et al. 2016 and refs. within). A submodule from the 0.2-2.7 µm eukaryotic fraction was also highly predictive of phosphate (Table 1, Fig. S8D). We identified the top four VIP taxa as an unclassified *Peronosporomycetes*, ...
an unclassified *Ochrophyta*, an unclassified Eukaryote, and an unclassified *Stramenopiles* (Fig. 6F, Table S1). All four OTUs had Pearson correlations with phosphate greater than 0.60, among which two were negative (Fig. 7A, Table S1). Unfortunately, the most highly interconnected OTU within the submodule remained unclassified even at the Phylum level (Fig. 7A, Table S1).

A > 2.7 \(\mu\)m eukaryotic submodule could predict 62% of measured phosphate variance (Table 1). An unclassified Eukaryote and an unclassified *Peronosporomycetes* occupied top two VIP positions (Fig. 7B, Table S1).

Only two submodules had strong associations with nitrate. The > 2.7 \(\mu\)m prokaryotic submodule most strongly correlated with nitrate contained 133 OTUs (Table 1, Fig. S9A). However, despite a low structural correlation to nitrate, the submodule strongly predicted measured nitrate in the PLS model (Figure S9B, Table 1). The four highest VIP scoring OTUs were an *Anabaena* sp., a *Flavobacterium* sp., an unclassified bacterium, and a member of the *Sphingobacteriales* NS11-12 marine group (Fig. 6B, Table S1). OTUs with the highest node centrality (> 20) belonged to the *Sphingomonadales* (*Alphaproteobacteria*) and *Sphingobacteriales* (Bacteroidetes), and each positively correlated with nitrate (\(r > 0.48\), Fig. 6B, Table S1). The > 2.7 \(\mu\)m eukaryotic submodule most associated with nitrate predicted 57% of observed variation in nitrate concentrations (Table 1). OTUs with the top VIP scores were two unclassified *Chrysophyceae*, an unclassified *Ochrophyta*, and an unclassified Diatom (Fig. 7C, S10; Table S1).

**Discussion**

Understanding microbial communities on a basin scale is essential for unveiling the complex linkage between organisms, metabolism, and water quality. Our near whole river 2914 km
transect revealed that while the MSR comprised similar aquatic microbial taxa found within other rivers (Ghai et al. 2011; Fortunato et al. 2013; Jackson et al. 2014; Read et al. 2015; Savio et al. 2015; Meziti et al. 2016), the river hosted unique relative abundance and core microbiome trends. Physio-chemical measurements demonstrated continual increases in nutrients and sediment along the upper river, whereas in the lower river, nitrate, phosphate, nitrite, and turbidity stayed more consistent (Fig. 1B). These distinct geographic patterns mirrored the observed separation of two microbial regimes at the Missouri River confluence (Fig. 2). This differed from the historical distinction of the upper and lower MSR at the Ohio River confluence in Cairo, IL, but matched the changes in Strahler order from eight to ten (Pierson et al. 2008). In general, both prokaryotic fractions and the > 2.7 µm fraction of eukaryotic communities increased in richness downriver (Fig. S3). The relative abundance of core community taxa decreased in the upper MSR before settling in the lower river (Fig. 5). Co-occurrence network analyses identified taxa associated with, and predictive of, the eutrophication nutrients nitrate and phosphate that provide important targets for future study and may assist in detecting and quantifying imminent changes in river water quality.

Some qualities of the MSR matched other rivers. MSR microbial communities separated into two distinct fractions of microbial assemblages, 0.2-2.7 µm and > 2.7 µm (ANOSIM, R = 0.65, P = 0.001), throughout the river transect (DeLong et al. 1993; Crump and Baross 1996; Allgaier and Grossart 2006; D’Ambrosio et al. 2014; Jackson et al. 2014). These fractions were composed of OTUs classified as LD12 (freshwater SAR11), Limnohabitans sp. (Betaproteobacteria), LD28 (Betaproteobacteria), acl clade (Actinobacteria), and Algoriphagus sp. (Bacteriodetes). These taxa also dominated in other rivers (Ghai et al. 2011; Fortunato et al. 2013; Kolmakova et al. 2014; Read et al. 2015; Savio et al. 2015; Meziti et al. 2016).
However, despite the MSR generally sharing similar taxa with other rivers, abundance patterns indicated a unique system. Phylum level abundances contrasted a previous study limited to the Minnesota portion of the MSR (Staley et al. 2013), as well as studies of the Thames (Read et al. 2015), Danube (Savio et al. 2015), and Columbia Rivers (Crump et al. 1999). Specifically, within the 0.2-2.7\(\mu\)m MSR prokaryotic fraction, Proteobacteria remained the most abundant phylum throughout the river, while Bacteroidetes and Actinobacteria decreased downriver (Fig. 3). In contrast, the other river studies identified Bacteroidetes-dominated headwaters that gave way to more abundant Actinobacteria and Proteobacteria further downriver (Crump et al. 1999; Staley et al. 2014a; Read et al. 2015; Savio et al. 2015). Although comparison beyond phylum level is important for understanding community changes, most studies report only OTU taxonomy making comparisons of OTU relative abundances difficult. OTUs belonging to the acl and LD12 freshwater clades continually increased in abundance towards the river mouth in the Thames and Danube Rivers (Read et al. 2015; Savio et al. 2015), whereas MSR OTUs classified as acl (OTUs 3, 4 and 7) and LD12 (OTU 11) did not change in abundance throughout the transect. Further, MSR OTU richness generally increased with distance, and core community abundance decreased along the upper river in both fractions, contrasting predictions by the RCC (Vannote et al., 1980) and observations in previous studies from other rivers (Read et al. 2015; Savio et al. 2015) where these trends were reversed. Importantly, we did not sample the true headwaters of the MSR (Lake Itasca to above St. Cloud), and at the point of first sampling, the MSR already constituted an eighth order river. Therefore, some of the trends predicted by the RCC may have been missed. Ultimately, the variant observations between this MSR survey and other river studies may result from different sampling methodologies or the particular timing of sampling. However, these differences may also stem from biological signal related to unique
environmental conditions, human impacts, and changes in the level of river engineering with distance.

Within the MSR, we observed distinct prokaryotic and eukaryotic microbial community regimes separated at the Missouri River confluence. This separation matched physio-chemical changes observed for the MSR such as increased Strahler river order (8 to 10) and more consistent nutrient concentrations (Fig. 1). We hypothesize that the separation between the upper and lower MSR communities was driven by changes in the importance of immigration, emigration, and resource gradient dynamics. Specifically, our data suggests that mass effects play a role in structuring microbial communities in the upper MSR, although instead of only in the headwaters, this process continues for almost a third of the length of the river. Increased turbidity correlated with decreases in freshwater associated bacteria (Spearman rank correlation, $> 2.7 \mu m r = 0.55$; $0.2-2.7 \mu m r = 0.51$) and the core microbiome relative abundance (Spearman rank correlation, $> 2.7 \mu m r = 0.53$; $0.2-2.7 \mu m r = 0.63$) in the upper MSR. Similarly, nutrients like phosphate and nitrate continually increased along the upper river (Fig. 1). In contrast, these nutrients, turbidity, and core microbiome relative abundance all stabilized in the lower MSR.

These patterns are consistent with communities under the influence of mass effects from tributaries in the upper MSR. However, once the MSR grew to a tenth order river, the large volume and size potentially buffered it from allochthonous influences, allowing species sorting effects to have greater influence over mass immigration. The lower river represents a more stable environment (e.g. nutrient concentrations) with its increased size and volume, contrasting the more variable upper MSR. Though river speed increased, the effective residence time also increased since taxa no longer experience rapidly changing environmental variables.
That we still observed increasing richness and variation between microbial communities along the lower MSR might indicate the increased influence of environmental filtering attributable to unmeasured bottom-up factors, such as the quality and quantity of DOM, or top-down influences such as predation or viral lysis. An overall community shift from a mixture of allochthonous members to a “native” population requires growth rates greater than residence time over a given distance (Crump et al. 2004). The extent to which a local or regional event impacts downriver populations would be dependent on the success of allochthonous taxa associated with those events to establish within the autochthonous population. The lower river provides ample opportunities for microbial community differentiation based on average prokaryotic growth rates (Savio et al. 2015), especially among particle-associated (> 2.7 µm) taxa (Crump et al. 1999). Thus, although the aggregate patterns in particular phyla and taxonomic richness within the MSR differ from other systems, similar ecological processes may still occur, but the relative proportion of the river whereby mass effects vs. species sorting dominate fosters unique community dynamics.

Microorganisms most associated with eutrophication nutrients are important target taxa for future study of nitrate and phosphate uptake and/or metabolizism, and also represent plausible biological indicators of river trophic state. By focusing on OTUs with the highest VIP scores in the PLS models, we identified several bacterial and eukaryotic taxa that fit these criteria (Figs. 6, 7). Organisms in the Sphingomonadaceae (e.g., Novosphingobium spp.) contributed strongly to the PLS models predicting nitrate with both > 2.7 µm (Fig. 6B) and 0.2-2.7 µm (Fig. S6A) size fractions (Table S1). Novosphingobium spp. isolates have previously been associated with eutrophic environments (Trusova and Gladyshev 2002; Zwart et al. 2002; Addison et al. 2007; Li et al. 2012) and some can reduce nitrate (Addison et al. 2007; Li et al. 2012), making
these specific OTUs candidates for nitrate metabolism in the river water column. An *Anabaena* sp. (OTU40) from the core microbiome had the top VIP score within the > 2.7 µm submodule that could predict 69% of nitrate concentrations, and correlated negatively with nitrate (Table 1, Fig. 6B). The nitrogen-fixing *Anabaena* (Allen and Arnon 1955) typically bloom in low dissolved inorganic nitrogen (DIN) conditions, making the absence of these consistent with high DIN. An unclassified *Holophagaceae* OTU (OTU33) had the highest VIP score within the prokaryotic 0.2-2.7 µm submodule that predicted 80% of the variance in phosphate concentrations (Table 1, Fig. 6A). This same OTU was also a key driver of the beta diversity separation between upper and lower river communities (Fig. 2B), and a member of the core microbiome (Table S1). The *Holophagaceae* belong to the Acidobacteria phylum, and the Ohio River has been reported to contain a much higher abundance of Acidobacteria relative to other tributaries (Jackson et al. 2014). Notably, Acidobacteria increased with river distance in our study as well (Fig. 3) with a peak in the 0.2-2.7 µm fraction around the Arkansas river. This increase in “free-living” Acidobacteria downriver is unique among other whole river studies, making these organisms, and the *Holophagaceae* OTUs in particular, potentially important organisms for the MSR river basin specifically.

Within Eukaryotes, multiple different algae, diatom, and Oomycetes OTUs occupied submodules highly predictive of nitrate and phosphate (Fig. 6A-C, Table S1), and specifically *Chrysophyceae* taxa from both size fractions correlated strongly with nitrate (Fig. 7C, Fig. S6E). *Chrysophyceae* (golden algae) commonly occupy river systems (Necchi Jr 2016) including the MSR (Korajkic et al. 2015), can be autotrophic and mixotrophic (Jansson et al. 1996), and may serve as predators of prokaryotes (Caron et al. 1990). Further, multiple OTUs classified as *Peronosporomycetes* were important in predicting phosphate in both size fractions (Fig 7A and
Peronosporomycetes are fungus-like eukaryotic organisms known to be pathogenic in fish, plants, and mammals (Dick 2003; Islam and von Tiedemann 2011). While we also identified many other eukaryotic OTUs as important predictors of nutrients, poor taxonomic resolution hindered our ability to discuss them further. Improved cultivation and systematics of key microbial eukaryotes will be vital to understanding river nutrient dynamics.

While the most geographically comprehensive analysis to date for the MSR, this study only encompasses a snapshot in time. Seasonal changes that have been observed in the Minnesota portion of the upper MSR and the Columbia River (Smith et al. 2010; Fortunato et al. 2013; Staley et al. 2015) undoubtedly influence this dynamic system. Future studies should incorporate microbial responses, at a whole-river scale, to seasonal pulse events (e.g. rain, snow melt) and how river size and volume may buffer local microbial communities from allochthonous inputs (Zeglin 2015). Our current research highlights the distinctiveness of MSR microbial communities and the complexities influencing their structure within the MSR ecosystem. While similar to other riverine studies in some respects, the 2014 MSR microbes differed in alpha diversity and core community membership compared to the other rivers. The observed association between changes in Strahler’s river order, nutrient dynamics, and community composition indicates the importance of hydrology (Freimann et al. 2015; Zeglin 2015; Niño-García et al. 2016) on the spatial dynamics structuring microbial communities and provides baseline information for future MSR studies that incorporate greater temporal and spatial resolution. With water quality and river health of growing local and global importance (Vorosmarty et al. 2010; Russell and Weller 2013), the determination of candidate taxa predictive of eutrophic nutrients provides impetus for targeted research on their larger potential in predicting, and potentially ameliorating, river quality.
Acknowledgements

This work was supported by the Department of Biological Sciences, College of Science, and the Office of Research and Economic Development at Louisiana State University; and the College of the Environment at University of Washington. Portions of this research were conducted with high performance computing resources provided by Louisiana State University (http://www.hpc.lsu.edu). The authors thank Dr. Gary King and Dr. Caroline Fortunato for thoughtful comments on the manuscript. The authors also thank the countless volunteers, schools, and organizations that facilitated the research. We specifically thank Pete Weess, Jessica Zimmerman, Katy Welch, Brian Moffitt, and David Cheney for helping organize the shipment of coolers between sites, and the OAR Northwest sponsors- Seattle Yacht Club Foundation, Yeti Coolers, and the National Mississippi River Museum and Aquarium. A full list of sponsors and volunteers can be found on the OAR Northwest website (oarnorthwest.org). We would also like to thank Dr. Matthew Sullivan and Dr. Simon Roux for their help with coding the WGCNA and sPLS analyses. Lastly, we would like to thank Mrs. Ginger Thrash, who connected the Thrash lab to OAR Northwest.

Conflict of Interest

The authors declare no competing financial interests.

References

nitrogenifigens sp. nov., a polyhydroxyalkanoate-accumulating diazotroph isolated from a

Allen, M. B., and D. I. Arnon. 1955. Studies on nitrogen-fixing blue-green algae. I. Growth and

Allgaier, M., and H.-P. Grossart. 2006. Seasonal dynamics and phylogenetic diversity of free-
living and particle-associated bacterial communities in four lakes in northeastern Germany.

Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable


are critical reservoirs of microbial diversity for fluvial networks. Proc. Biol. Sci. 280:
20131760.

Allison. 2010. The science of hypoxia in the northern Gulf of Mexico: A review. Sci. Total
Environ. 408: 1471–1484.

Blanchet, M., O. Pringault, C. Panagiotopoulos, D. Lefèvre, B. Charrière, J.-F. Ghiglione, C.
When riverine dissolved organic matter (DOM) meets labile DOM in coastal waters:
changes in bacterial community activity and composition. Aquat. Sci. , doi:10.1007/s00027-


Dick, M. 2003. Straminipilous Fungi: Systematics of the Peronosporomycetes Including Accounts of the Marine Straminipilous Protists, the Plasmodiophorids and Similar

Duan, S., R. T. Powell, and T. S. Bianchi. 2014. High frequency measurement of nitrate concentration in the Lower Mississippi River, USA. J. Hydrol. 519: 376–386.


Wood, S. A., M. J. Prentice, K. Smith, and D. P. Hamilton. 2010. Low dissolved inorganic nitrogen and increased heterocyte frequency: precursors to Anabaena planktonica blooms in
a temperate, eutrophic reservoir. J. Plankton Res. 32: 1315–1325.


Table 1. Correlations between prokaryotic and eukaryotic submodules to nitrate and phosphate, as well as PLS model results

<table>
<thead>
<tr>
<th>Fraction</th>
<th>Gene</th>
<th>Nutrient</th>
<th>Submodule</th>
<th>Eigen Correlations</th>
<th>Total OTUs</th>
<th>PLS model</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.2-2.7 µm</td>
<td>16S</td>
<td>NO₃⁻</td>
<td>Blue</td>
<td>r = 0.6, P = 7e-08</td>
<td>77</td>
<td>R²=0.35; corr = 0.65, P =1.e-09</td>
</tr>
<tr>
<td>&gt; 2.7 µm</td>
<td>16S</td>
<td>NO₃⁻</td>
<td>Brown</td>
<td>r = 0.56, P = 3e-07</td>
<td>133</td>
<td>R² = 0.69; corr = 0.83, P = &lt; 2.2e-16</td>
</tr>
<tr>
<td>0.2-2.7 µm</td>
<td>16S</td>
<td>PO₄³⁻</td>
<td>Turquoise</td>
<td>r = 0.53, P = 2e-06</td>
<td>51</td>
<td>R²=0.80; corr = 0.89, P = &lt; 2.3e-16</td>
</tr>
<tr>
<td>&gt; 2.7 µm</td>
<td>16S</td>
<td>PO₄³⁻</td>
<td>Yellow</td>
<td>r = 0.53, P = 1e-06</td>
<td>80</td>
<td>R²= 0.48; corr= 0.77, P = 4.88e-15</td>
</tr>
<tr>
<td>0.2-2.7 µm</td>
<td>18S</td>
<td>NO₃⁻</td>
<td>Green</td>
<td>r = 0.39, P = 6e-04</td>
<td>39</td>
<td>R²=0.38; corr = 0.62, P =2.97e-9</td>
</tr>
<tr>
<td>&gt; 2.7 µm</td>
<td>18S</td>
<td>NO₃⁻</td>
<td>Brown</td>
<td>r = 0.52, P = 2e-06</td>
<td>59</td>
<td>R²=0.572; corr = 0.759, P = 6.7e-15</td>
</tr>
<tr>
<td>0.2-2.7 µm</td>
<td>18S</td>
<td>PO₄³⁻</td>
<td>Turquoise</td>
<td>r = 0.56, P = 2e-07</td>
<td>56</td>
<td>R²=0.80; corr = 0.89, P = &lt; 2e-16</td>
</tr>
<tr>
<td>&gt; 2.7 µm</td>
<td>18S</td>
<td>PO₄³⁻</td>
<td>Red</td>
<td>r = 0.60, P = 2e-08</td>
<td>39</td>
<td>R²= 0.618; corr = 0.799, P = &lt; 2e-16</td>
</tr>
</tbody>
</table>
**Figure Legends**

**Figure 1.** Sampling map (A) with graph inserts that represent the measured discharge rate (cubic feet second$^{-1}$ [CFS]) as recorded on the USGS gauges, and six environmental parameters measured along the transect, according to concentration or visible depth of secchi disk by distance (B). Throughout the figure and text, blue and red dots represent sampling locations above and below the Missouri River confluence, respectively, and are designated throughout as “Upper” and “Lower.” Cell Counts only represent the < 2.7 µm fraction.

**Figure 2.** Non-metric multidimensional scaling (NMDS) results for whole community with correlations to environmental parameters (phosphate, nitrate, nitrite, ammonia, distance (km), water temperature, turbidity (cm), river speed (mph)) and the top ten OTUs based on significance (P) and strength of correlation (r). The four plots represent > 2.7 µm (A, C) and 0.2-2.7 µm (B, D) fractions for the 16S (A, B) and 18S (C, D) rRNA gene communities. Vector length is proportional to the strength of the correlation.

**Figure 3.** Relative abundance, by phylum, according to transect distance, for phyla accounting for > 0.1% of the total reads for the 16S rRNA gene > 2.7 µm (A) and 0.2-2.7 µm (B) communities. Non-linear regressions with 95% CI (gray shading) are provided for reference.

**Figure 4.** Relative abundance, by phylum, according to transect distance, for phyla accounting for > 0.1% of the total reads for the 18S rRNA gene > 2.7 µm (A) and 0.2-2.7 µm (B) communities. Non-linear regressions with 95% CI (gray shading) are provided for reference.

**Figure 5.** Core microbiome aggregate abundance for the 16S (A) and 18S (B) rRNA gene. In each, triangles and circles points represent 0.2-2.7 µm and > 2.7 µm fractions,
respectively. Non-linear regressions with 95% confidence intervals (CI) (gray shading) are provided for reference.

**Figure 6.** PLS results for the 16S rRNA gene community submodules most associated with phosphate (A) and nitrate (B). OTU correlation with a given nutrient is indicated on the y-axis according to the number of co-correlations (node centrality) on the x-axis. Community fractions: 0.2-2.7 µm (A) and > 2.7 µm (B). Circle size is proportional to VIP scores, with top 10 VIP scoring and top node centrality OTUs labeled with their highest-resolution taxonomic classification and OTU number. Colors represent the taxonomic classification the phylum level.

**Figure 7.** PLS results for the 18S rRNA gene community submodules most associated with phosphate and nitrate. OTU correlation with a given nutrient is indicated on the y-axis according to the number of co-correlations (node centrality) on the x-axis. Community fractions: 0.2-2.7 µm (A) and > 2.7 µm (B,C). Circle size is proportional to VIP scores, with top 10 VIP scoring and top node centrality OTUs labeled with their highest-resolution taxonomic classification and OTU number. Colors represent the taxonomic classification the phylum level.
Supplemental Figure Legends

**Figure S1.** NMDS results for the 16S (A) and 18S (B) rRNA gene communities. In each, circles and triangles represent the > 2.7 µm and 0.2-2.7 µm fractions, respectively.

**Figure S2.** Hclust results for the 16S (A) and 18S (B) rRNA gene communities.

**Figure S3.** Richness and Evenness indexes for the 16S > 2.7 µm (A) and 0.2-2.7 µm (B) and 18S > 2.7 (C) and 0.2-2.7 µm (D) communities. Non-linear regressions with 95% CI (gray shading) are provided for reference.

**Figure S4.** Separated upper and lower MSR core microbiome aggregate abundance for the 16S rRNA gene communities. For the upper and lower river, the core microbiome was defined separately requiring OTUs to have greater than one read in 90% of the samples. > 2.7 µm (A) and 0.2-2.7 µm (B) 16S rRNA gene communities in the upper and lower MSR.

**Figure S5.** Relative abundance, by environmental ontology group, according to transect distance for the 16S rRNA gene > 2.7 µm (A) and 0.2-2.7 µm (B) communities. Non-linear regressions with 95% CI (gray shading) are provided for reference.

**Figure S6.** PLS results for the 0.2-2.7 µm and > 2.7 µm 16S (A-D) and 18S (E-H) rRNA gene community for selected submodules with nitrate and phosphate and a VIP score > 1. Correlation of submodule OTUs to nitrate and phosphate according to the number of co-correlations (node centrality) for 0.2-2.7 µm (A,B, E,F) and > 2.7 µm (C,D,G,H) 16S and 18S rRNA gene communities. Circle size is proportional to VIP scores, with top 10 VIP scoring and top node centrality OTUs labeled with their highest-resolution taxonomic classification and OTU number. Colors represent the taxonomic classification the phylum level.

**Figure S7.** WCGNA results for 0.2-2.7 µm 16S rRNA gene community submodules of interest
based on Pearson correlations to nitrate and phosphate (A, B, D) and the resultant PLS regression of the predicted versus measured nutrient concentrations (C, E). Linear regressions with 95% CI (gray shading) are provided for reference.

**Figure S8.** WCGNA results for 0.2-2.7 µm 18S rRNA gene community submodules of interest based on Pearson correlations to nitrate and phosphate (A, B, D) and the resultant PLS regression of the predicted versus measured nutrient concentrations (C, E). Linear regressions with 95% CI (gray shading) are provided for reference.

**Figure S9.** WCGNA results for > 2.7 µm 16S rRNA gene community submodules of interest based on Pearson correlations to nitrate and phosphate (A, B, D) and the resultant PLS regression of the predicted versus measured nutrient concentrations (C, E). Linear regressions with 95% CI (gray shading) are provided for reference.

**Figure S10.** WCGNA results for > 2.7 µm 18S rRNA gene community submodules of interest based on Pearson correlations to nitrate and phosphate (A, B, D) and the resultant PLS regression of the predicted versus measured nutrient concentrations (C, E). Linear regressions with 95% CI (gray shading) are provided for reference.
**Supplemental Tables**

Supplemental Table S1 is a spreadsheet, TableS1.xlsx. This includes site data, NMDS correlations for environmental data, Relative abundances of the Core taxa, SeqENV data, 16S and 18S rRNA OTU tables, and VIP scores for all samples and datum.

Additional Supplemental Information, including R scripts and our Mothur workflow, are hosted on the Thrash Lab website at: thethrashlab.com/publications.
Figure 2

The copyright holder for this preprint (which was not peer-reviewed) is the author/funder. It is made available under a CC-BY-NC-ND 4.0 International license.
Figure 3

A

B