Microbial regime changes and indicators of eutrophication on the Mississippi River identified via a human-powered 2900 km transect

Michael W. Henson[1], Jordan Hanssen[2], Greg Spooner[2], Patrick Fleming[2], Markus Pukonen[2], Frederick Stahr[3], and J. Cameron Thrash[1*]

[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: thrashc@lsu.edu

Abstract

Draining 31 states and roughly 3 million km^2, 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 20^{th} 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 16S and 18S rRNA gene amplicons and physicochemical samples from 38 MSR sites over nearly 3000 km from Minnesota to the Gulf of Mexico. These data represent the first of their kind for a major river and revealed distinct regime changes between upper and lower MSR microbial communities that corresponded to Strahler's river order and nutrient concentrations. Within these assemblages, we identified subgroups of OTUs from the phyla Acidobacteria, Bacteroidetes, and Heterokonts that highly correlated with, and were predictive of, important eutrophication nutrients. This study offers the most comprehensive view of Mississippi River microbiota to date, establishes the groundwork for future temporal and spatial studies of river perturbations, and provides potential microbial indicators of river health.

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[1–5]. Within the carbon cycle alone, rivers collectively discharge organic carbon to the oceans at over 0.4 Pg C yr^{-1} [6]. Perhaps more importantly, rivers are generally net heterotrophic[7], 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} [5,8]. 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, the fundamental engines of these transformations- microorganisms- have received comparatively little study in rivers relative to other aquatic systems, despite the fact that microbial functions likely play a vital role in ecosystem health for both rivers themselves and their places of discharge.

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[9,10] 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)[11], but can exceed 3 million cfs during flood stages[12], carrying over 150 x 10^9 kg of suspended sediment into the northern GOM annually[9,13]. This massive discharge includes excess nutrients (nitrogen and phosphorus), primarily from agricultural runoff [14–18], and fuels one of the largest marine zones of seasonal hypoxia in the world[19–22]. Understanding microbial relationships to river eutrophication will inform hypotheses regarding 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[23–25], attributable to changing nutrient concentrations[26–28], dendritic length[4,29], differing dissolved organic matter (DOM) sources[26,30,31], and land use changes[16,27,28,32]. Past studies of the Thames, Danube, Yenisei, and Columbia Rivers have found that planktonic river microbiota were dominated by the phyla Actinobacteria, Proteobacteria, and Bacteriodetes, specifically, taxa such as- hgcI/cal Actinobacteria, *Polynucleobacter* spp., GKS9, and LD28 *Betaproteobacteria*, CL500-29 Bacteriodetes, LD12 freshwater SAR11 *Alphaproteobacteria*, and *Novosphingobium* spp.[4,33,34]. 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[4] or richness[29] of freshwater taxa with river distance, and an increased abundance of "core" river taxa[35] with cumulative residence time[44,22,27,33, 52].

Researchers have suggested that these patterns supported application of the River Continuum Concept (RCC)[10] to river microbiota. The concept 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[36], and richness will increase from headwaters to mid-order before decreasing in higher order rivers[36]. 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[4,37,38]. Complicating matters, particle-associated communities in rivers (frequently defined as those found on filters of > ~3 μm) remain distinct from their free-living counterparts[37–39, 53], potentially due to increased production rates from readily obtainable carbon[38, 53]. Typical taxa associated with particles include OTUs related to the Bacteroidetes clades *Flavobacteria* and *Cytophaga*, Planctomycetes, *Rhizobium* spp., and *Methylophilaceae* spp[39,42–44]. 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[42,45]. Thus, rivers constitute complex and highly dynamic ecosystems from a metacommunity perspective.

However, our knowledge of river microbial assemblages, their ecology, dispersal, and their relationship to chemical constituents remains in comparative infancy to that of lakes and oceans. Although microbes play a central role in the RCC[36], microbiologically-oriented transects at the whole-river or -catchment scale have only been conducted for a handful of systems[4,28,29,46,47], and until this work, none had been attempted for any of the largest rivers in the world. Furthermore, little data exists on microbial eukaryotic communities in rivers. 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. Rowers from the adventure education non-profit OAR Northwest collected samples from Minneapolis, MN to the Gulf of Mexico (2918 km) (Fig. 1A). They also maintained active blogging and social media content and visited 21 schools along the river that incorporated elements of the journey into their curriculum. Our findings greatly expand the current information available on microbial assemblages in major lotic ecosystems and help further delineate the relationships between microbial structure and stream order, nutrients, and volume.

Results

Using rowboats and 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[48]), which corresponded to overall discharge (Fig. 1A) and beta diversity changes discussed below, and thus we used this juncture to separate the upper MSR (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). Ammonium showed much greater variability along the transect. 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 in the upper MSR, and decreased to high 10^5 cells/mL 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 values that trended upwards downriver (Figs. S3A-B), although an earlier peak occurred for both at sites O-Q (761-999 km) below the Des Moines River and above the Illinois River (Figs. S3A-B). Both size fractions (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) (Figs. 2A and B), that the Eukaryotic fractions mirrored (below). The environmental variables phosphate and turbidity had the highest correlation with the separation between 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 hgcI/cal 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 most strongly correlated with the distinction between upper and lower communities among environmental factors (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 a 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 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 showed wide fluctuations 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, but showed considerable variation in abundance in the larger fraction. Cyanobacteria in the > 2.7 μm fraction strongly and negatively correlated with increased turbidity (Spearman rank = 0.67). Both > 2.7 μm and 0.2-2.7 μm Acidobacteria had a strong positive correlation with river distance (Wilcoxon single ranked test, P = < 0.01) (Fig. 3A and B). Within the 0.2-2.7 μm fraction, the five most abundant OTUs were classified as LD12 (OTU11), two hgcI clade OTUs (OTU4, OTU7), *Limnohabitans* sp. (OTU2), and LD28 (OTU8) (Table S1). Comparatively, an unclassified *Methylophilaceae* (OTU1), *Planktothrix* sp. (OTU12), NS11-12 marine group (OTU21), *Aquirestis* sp. (OTU17), and unclassified *Sphingobacteriales* (OTU25) were the most abundant OTUs in the > 2.7 μm (Table S1). Archaea occurred at much lower relative abundances: we found only eight OTUs belonging to the Euryarchaeota and Thaumarchaeota. 0.2-2.7 μm Thaumarchaetoa increased in abundance along the transect (Fig. 3B), but at less so than those in the > 2.7 μm fraction (Fig. 3A). In both fractions, we only detected Euryarcheota at very low abundances.

We defined the core microbiome as those OTUs detectable after normalization in greater than 90% of the samples. The > 2.7 μm and 0.2-2.7 μm core microbiomes consisted of 95 and 106 OTUs, respectively, classified into eight different phyla- Proteobacteria, Actinobacteria, Bacteroidetes, Cyanobacteria, Verrucomicrobia, Chloroflexi, Chlorobi, Gemmatimonadetes- and some remained unclassified (Table S1). Core microbome relative abundance in both fractions decreased along the upper river but stabilized in the lower river (Fig. 4A). 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).

16S rRNA gene environmental ontology

We successfully classified 313 of the 945 OTUs with EnvO terminology (Table S1). Of those, freshwater organisms dominated, although their relative abundance in both the > 2.7 μm and 0.2-2.7 μm fractions decreased with river distance before stabilizing in the lower MSR (Fig. S5A-B). However LD12, the most abundant OTU in our dataset and a well-established freshwater organism, did not receive an EnvO classification at all, indicating the limitations of this technique with current database annotations. Terrestrial organisms from the > 2.7 μm fraction decreased along the transect (Fig. S5A), while the 0.2-2.7 μm fraction remained stable (Fig. S5B). Although representing a minor fraction of

total OTUs, sediment-associated microorganisms remained steady along the river in both fractions.. Taxa associated with anthropogenic sources in both fractions increased along the river(Fig. S5A-B).

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 (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 among 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 lower and upper 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 constituted 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 unclassified Eukaryote (OTU1) separated the MSR communities in the 0.2-2.7 μm (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. 5). 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, which include fungi (Fig. 5A; Table S1). Within the 0.2-2.7 μm fraction, we classified the five most abundant OTUs as three unclassified *Bacillariophytina* (OTU7, OTU14, OTU9), a *Pythium* sp. (OTU170), and a unclassified *Cryptomonas* (OTU11) (Table S1). Comparatively, two unclassified *Eukaryotes* (OTU2 and OTU1), a unclassified *Stramenopiles* (OTU3), a unclassified *Perkinsidae* (OTU13), and a unclassified *Chrysophyceae* (OTU6) had the highest abundance in the > 2.7 μm (Table S1).

Eighty OTUs comprised the > 2.7 μm core microbiome, averaging 28% of aggregate community relative abundance (Fig. 4B). We classified these as *Alveolata*, *Cryptophceae*, *Nucletmycea*, *Stramenopiles*, or unclassified Eurkaryota (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 that, in aggregate, averaged 20% of the community relative abundance across all samples (Fig. 4B). These OTUs consisted of *Alveolata*, *Nucletmycea*, *Stramenopiles*, or unclassified Eurkaryota (Table S1). While the 0.2-2.7 μm core

284 microbiome remained relatively stable along the river, the > 2.7 μm core
285 decreased along the upper MSR before stabilizing in the lower river, similarly to
286 that of the prokaryotes (Fig. 4B).
287
288 *Network analyses identify indicator taxa associated with eutrophication*
289 Since increased nitrogen and phosphorous is of interest for stakeholders both
290 along the river and in the GOM, we evaluated relationships between individual
291 OTUs and these nutrients to identify taxa indicative of eutrophication. Among 0.2-
292 2.7 μm prokaryotes, a co-correlation network (submodule) strongly associated
293 with nitrate (r = 0.6, P = 7e-08) (Fig. S6A) comprised 77 OTUs, mostly from the
294 Proteobacteria. OTU relative abundances explained 42% of the variance in
295 nitrate (LOOCV, R^2=0.65; corr = 0.65, P =1.e-09), and the top four OTUs by VIP
296 score belonged to the Proteobacteria and Bacteroidetes (Fig. 6B; FigS6A; Table
297 S1). Of these, three were highly correlated to nitrate (r > 0.50) (Table S1):
298 unclassified *Comamonadaceae, Mucilaginibacter,* and *Pseudospirillum* spp..
299 OTUs that had more than 8 node connections, indicating taxa co-correlating with
300 numerous others, included a *Mucilaginibacter* sp., a *Novosphingobium* sp., and a
301 *Caulobacter* sp. (Fig. 6B, Table S1).
302 The submodule best correlated with phosphate (r = 0.53, P = 2e-06) (Fig.
303 S6) comprised 151 OTUs that explained 80% of the variance (LOOCV, R^2=0.80;
304 corr = 0.89, P = < 2.3e-16) (Fig. S6D). OTUs with VIP scores > 1 belonged to
305 seven different phyla (Table S1), with the top four identified as an unclassified
306 *Holophagaceae*, an unclassified *Gemmatimonadaceae*, an unclassified
307 *Burkholderiaceae,* and a *Pseudosprillum* sp. (Table S1). Three of these four
308 OTUs had Pearson correlations with phosphate greater than 0.62 (Fig. 6B). The
309 most highly interconnected OTU within the submodule was an unclassified
310 Acidobacteria (Table S1).
311 In the > 2.7 μm fraction, the prokaryotic submodule with the highest
312 correlation to nitrate (r = 0.56, P = 3e-07) (Fig. S7A) had 133 OTUs and
313 explained 69% of the variation in nitrate (LOOVC, R^2 = 0.69; corr = 0.83, P= <
314 2.2e-16). The four highest VIP scoring OTUs, an *Anabaena* sp., a
315 *Flavobacterium* sp., an unclassified bacterium, and a member of the
316 *Sphingobacteriales* NS11-12 marine group, anticorrelated with nitrate (Fig. 7A,
317 Table S1). A NCBI BLAST of the unclassified bacterium returned no significant
318 hits > 90% identity to named organisms. OTUs with the highest node centrality
319 (> 20) belonged to two clades of bacteria, *Sphingomonadales*
320 (*Alphaproteobacteria*) and *Sphingobacteriales* (Bacteroidetes), and all correlated
321 positively to nitrate (r > 0.48, Table S1).
322 When considering phosphate, the most significant submodule showed a
323 modest correlation (r = 0.53, P = 1e-06, Fig. S7A), but OTU abundances could
324 only explain 48% of the variation of measured phosphate (LOOVC, R^2= 0.48;
325 corr= 0.77, P = 4.88e-15). Proteobacteria constituted the majority of the 80 OTUs
326 from the submodule, and the top four scoring VIP OTUs were an unclassified
327 *Gammaproteobacteria*, an *Arcicella* sp., an unclassified *Cytophagaceae*, and a

Nitrospira sp. (Fig. 7B, Table S1); the latter two had moderate correlations to phosphate (r = > 0.55). The OTUs with the highest number of node connections were a *Woodsholea* sp. from the *Caulobacterales* family and the same *Nitrospira* OTU (Table S1).

Among 0.2-2.7 μm Eukaryotic size fraction, a submodule strongly associated with nitrate (r = 0.39, P = 6e-04) (Fig. S8A) comprised 39 OTUs, mostly from the phylum Stramenopiles. OTU relative abundances explained 38% of the variance in nitrate (LOOCV, R^2=0.38; corr = 0.62, P =2.97e-9), and the top four OTUs by VIP score classified as an unclassified *Chrysophyceae*, an unclassified *Ochrophyta*, and an unclassified *Chromulinales* (Fig. 8B; FigS8A; Table S1). Of these, two highly correlated to nitrate (r > 0.49) (Table S1): the *Chrysophyceae,* and *Chromulinales* OTUs.

The submodule best correlated with phosphate (r = 0.56, P = 2e-07) (Fig. S8A) comprised 56 OTUs that explained 80% of the variance (LOOCV, R^2=0.80; corr = 0.89, P = < 2e-16). OTUs with VIP scores > 1 belonged to four different phyla (Table S1), with the top four identified as an unclassified *Peronosporomycetes*, an unclassified *Ochrophyta*, an unclassified Eukaryote, and an unclassified Stramenopiles. (Table S1). All four OTUs had Pearson correlations with phosphate greater than 0.60, two were negative (Fig. 8B). The most highly interconnected OTU within the submodule was an unclassified Eukaryote (Table S1).

Among the > 2.7 μm Eukaryote taxa, submodules with strongest correlations to nitrate and phosphate (NO_3^-: r = 0.52, P = 2e-06; PO_4^{3-}: r = 0.60, P = 2e-08, Fig. S9A) had smaller membership than those for prokaryotes, comprising 59 and 39 OTUs, respectively. The OTUs in the submodule most strongly correlated with nitrate could predict 57% of observed variation in nitrate (LOOVC, R^2=0.572; corr = 0.759, P= 6.7e-15). OTUs with top VIP scores were two unclassified Chrysophyceae, an unclassified Ochrophyta, and an unclassified Diatom (Fig. 8B; Table S1). When considering phosphate, submodule OTU abundances predicted 62% of measured concentrations (LOOVC, R^2 = 0.618; corr = 0.799, P = < 2e-16). An unclassified Eukaryote and an unclassified Peronosporomycetes occupied top two phosphate-associated positions according to VIP score (Fig. 9B; Table S1).

Discussion

Our 2914 km transect of the MSR has supplied the longest microbiologically oriented transect of a top ten river based on volume, length, or drainage. Water samples illustrated a river continually inundated with nutrients and sediment in its upper portion that gave way to more consistent levels in the lower river (Fig. 1B). These distinct scenarios mirrored the different microbial regimes separated by the Missouri river confluence (Fig. 2), which differed from the historical distinction of the upper and lower MSR at the Ohio River confluence in Cairo, IL, but matched the separation based on changes in Strahler order from eight to ten[48]. In general, both prokaryotic fractions and the > 2.7 μm fraction of

eukaryotic communities increased in richness downriver while the percent of core community taxa decreased in the upper MSR before settling in the lower river (Figs. 4, S3), concomitant with increased stability of environmental factors (Fig. 1B). Co-occurrence network analyses identified important potential indicator taxa for the eutrophication nutrients nitrate and phosphate that may help future efforts to detect and quantify imminent changes in river water quality.

In general, the most abundant OTUs throughout the MSR corresponded to typical freshwater taxa observed in other important riverine/aquatic studies[4,25,29,35,49,50], such as LD12 (*Alphaproteobacteria*), hgcI-cal clade (Actinobacteria), *Polynucelobacter* (*Betaproteobacteria*), LD28 (*Betaproteobacteria*), and *Limnohabitans* (*Betaproteobacteria*) (Table S1). Though our nutrient measurements mirrored conditions found in previous studies of the Danube, Thames, and upper Mississippi rivers[4,27,34], the relative abundances of important phyla differed. Specifically, in our study Proteobacteria remained the most abundant phylum, while Bacteroidetes and Actinobacteria decreased (Fig. 3), whereas in the Minnesota portion of the MSR[35], Thames[29], Danube[4], and Columbia Rivers[40], Bacteroidetes dominated headwaters while Actinobacteria dominated further downriver. General increases in MSR species richness with distance and decreases in the percent core community along the upper river contrasted predictions by the RCC[36] and observations in previous studies[4,27] where these trends were reversed. Importantly, we did not sample the true headwaters of the MSR (Lake Itasca to above St. Cloud), and therefore at the point of first sampling, the MSR already constituted an eighth order river. Ultimately, some of these variant observations may result from different sampling methodologies, but also from biological signal related to unique environmental conditions and human impacts, changes in the level of river engineering with distance, and the magnitude of the MSR (in terms of volume and catchment complexity) and its tributaries relative to previously sampled systems. However, perhaps some differences lie in the proportion and scale at which the varied influences of dispersal and environmental filtering occur.

Many of our results suggest similar underlying ecological mechanisms with other river systems, even though the specific microbial community patterns differ. Multiple studies have shown evidence for the importance of mass effects in headwaters, while species sorting dominates with increased residence time as rivers grow in size[44,52]. Our data suggests that mass effects indeed play a role 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 bacteria (Spearman rank correlation, > 2.7 μm R = 0.55; 0.2-2.7 μm R = 0.51) and the core microbiome (Spearman rank correlation, >2.7 μm R = 0.53; 0.2-2.7 μm R = 0.63) during the first ~1000 km- the upper MSR- whereas these variables and some nutrient concentrations (Fig. 1) stabilized in the lower MSR. These patterns are consistent with communities under the influence of mass effects from tributaries in the upper MSR. 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 dominate. The lower river represents a more stable environment (e.g. nutrient concentrations) with its increased size and volume, contrasting the more variant upper MSR. Though the river speed increases, the effective residence time also increases since taxa no longer experience rapidly changing environmental variables.

That we still observed variation between microbial communities along the lower MSR concurs with 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 that allow taxa to overcome mass effects over a given distance[53]. The lower river also provides ample opportunities for microbial community differentiation based on average prokaryotic growth rates[4], especially among particle-associated (> 2.7 μm) taxa[40]. Thus, while the aggregate patterns in particular phyla and taxonomic richness may differ from other systems, similar ecological processes may still occur, but the relative proportion of distance whereby mass effects vs. species sorting dominate fosters unique community dynamics.

Within microbial communities lie taxa with unique roles as potential mitigators of human impact. MSR water quality has continually been rated low[54], suffers from significant eutrophication, and causes one of the largest worldwide zones of seasonal hypoxia in the northern GOM[21]. Identification of microorganisms with relationships to nitrogen and phosphorous will inform efforts to model nutrient remediation and provide important targets for future study. However, such microbial indicators of biological integrity (IBIs- metrics to quantify the health of an aquatic system[55]), require better development[55–57]. Phosphate and nitrate increased along the upper MSR and generally stabilized in the lower river (Fig. 1). Using co-correlation networks and partial least squares modeling (PLS), we identified submodules with containing taxa with significant predictive power for phosphate and nitrate concentrations (Figs. 6-9). Individual OTUs within these submodules with strong correlations to nitrate or phosphate, and VIP scores > 1 within the PLS models, represent potential indicator taxa for these nutrients.

Taxa in a 0.2-2.7 μm fraction submoule could predict 80% of the variance in phosphate concentrations, with much weaker correlations to nitrate. A *Holophagaceae* OTU (OTU33) was the top 0.2-2.7 μm taxon for predicting phosphate concentrations based on VIP scores (Fig. 6B). It also occupied an important role in separating communities at a beta-diversity level (Fig. 2B), and resided in the core microbiome. A 2014 study on the tributaries of the MSR found that the Ohio River had much higher abundance of Acidobacteria relative to other tributaries[39], which corroborates our finding of increasing Acidobacteria downriver. Although the *Holophagaceae* comprises diverse and abundant taxa[58], many environmentally relevant taxa remain elusive to cultivation efforts[59]. One of the few cultivated representatives from the *Holophagaceae* family, *Holophaga foetida,* has the genomic capacity to accumulate phosphorous and synthesize

polyphosphate via a polyphosphate kinase[60]. While this organism has an obligately anaerobic lifestyle* that makes it an unlikely match for the OTU, it indicates genomic potential that may span the clade. An unclassified *Gemmatimonadaceae* (OTU60) from the 0.2-2.7 μm fraction correlated strongly with phosphate and also played a strong role in driving beta-diversity changes between the upper and lower river (Fig. 2)

Contrasting the 0.2-2.7 μm taxa, a submodule in the > 2.7 μm fraction could predict 69% of nitrate concentration variance. Within the prokaryotic > 2.7 μm taxa, an *Anabaena* sp. (OTU40) had the top VIP score among submodule taxa predicting nitrate, correlated negatively (Fig. 7A), and had membership in the core microbiome. The nitrogen-fixing *Anabaena*[61] typically bloom in low dissolved inorganic nitrogen (DIN) conditions, making the absence of these consistent with high DIN. Organisms in the *Sphingomonadaceae* (e.g., *Novosphingobium* spp.) also contributed strongly to the PLS models predicting nitrate with both > 2.7 μm (Fig. 7A) and 0.2-2.7 μm (Fig. 6A) taxa (Table S1). *Novosphingobium* isolates have previously been associated with eutrophic environments[34,64–66] and some can reduce nitrate[64,65], making these OTUs good candidate IBIs within the MSR basin.

Notably, a *Nitrospira* sp. (OTU96) significantly correlated with phosphate (R^2=0.60, P=<0.001) and and was the second most important taxa for predicting phosphate concentrations in the > 2.7 μm fraction (VIP = 1.8). *Nitrospira* occupy a key role as nitrite oxidizers, and have been found in various environments including the “Dead Zone” in the GoM[22] and in correlation with wastewater treatment effluent[67]. This OTU makes an intriguing IBI candidate in a watershed impacted by wastewater[68] and that directly influences the Dead Zone. Additionally, as an organism that creates nitrate, this might represent a taxon that exacerbates, rather than mitigates, the effects of eutrophication.

Within Eukaryotes, multiple algae and diatoms strongly correlated with nitrate and phosphate (Fig. 8 and 9; Table S1), and specifically *Chrysophyceae* taxa from both fractions correlated strongly with nitrate (Fig. 8A). *Chrysophyceae* (golden algae) commonly occupy river systems[69] including the MSR[68], can be autotrophic and mixotrophic[70], and may serve as predators of prokaryotes[71]. While we also identified many other eukaryotes 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 to date, this study only encompasses a snapshot in time for the MSR. Seasonal changes that have been observed in the Minnesota portion of the upper MSR[52] and the Columbia River[52,72,73] undoubtedly influence this dynamic system. Future studies should incorporate microbial responses, at a full 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. Our current research highlights the uniqueness and complexities of the MSR ecosystem. The observed association

between changes in Strahler's river order, nutrient dynamics, and community composition indicates the importance of hydrology[46,47] 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[54,74], the determination of candidate microbial IBIs also provides impetus for targeted research on their functions and further investigation of these organisms as sentinels of river health.

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 25mm 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-AI 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[75,76] 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[77].

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[78], and eukaryotic sequences from the V9 region of the 18S rRNA gene using the 1391r and EukBR primer set[79]. 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 13253140 and 13240531 sequences, respectively.

Sequence Analysis

We analyzed amplicon data with Mothur v.1.33.3[80] using the Silva v.119 database[81,82]. 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[83], 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 into operational taxonomic units (OTUs) using a 0.03 dissimilarity threshold ($OTU_{0.03}$). After these steps, 146725 and 131352 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[84]. 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. 146725 and 131327 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[85]. Using the package *PhyloSeq*[86], alpha-diversity was first calculated on the unfiltered OTUs using the "estimate richness" command within PhyloSeq, which calculates Chao1[86]. 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[87] with a variance stabilizing transformation[88]. 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 of the NMDS analyses using the *anosim* function in the vegan package[84]. The influence of environmental parameters on beta-diversity was calculated in R with the *envfit* function.

Network analyses and modeling

To identify specific OTUs with strong relationships to environmental parameters (e.g. turbidity, NO_3^-), we employed weighted gene co-expression network analysis (WGCNA)[89] as previously described[90] for OTU relative abundances. 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 μm, p = 4 for 16S 0.2-2.7 μm) 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. S6-8 A). To explore the relationship of submodule structure to these parameters, submodule OTUs were plotted using their individual correlation to said parameter (here nitrate or phosphate) and their submodule membership, defined as the number of connections within the module (Figs. S6-8 B, 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) regression 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 permutated a 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. S6-8 C, E). Relative contributions of individual OTUs to the PLS regression were calculated using value of importance in the projection (VIP)[91] determination. PLS was run using the R package *pls*[92], while VIP was run using an 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) as previously described[4]. 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 eight groups: Freshwater, Aquatic Undetermined, Salt Water, Anthropogenic, Other, 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 equal distribution between two or more groups were classified as Unclassified. OTUs that returned no significant hits to an EnvO term were assigned to a ninth category, NA. OTUs and their corresponding relative abundances were merged based on the assigned group and plotted.

Accession numbers

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

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 “Supplementary Data”.

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. We would like to thank Dr. Gary King and Dr. Caroline Fortunato for their friendly reviews. 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, Yetti 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 support and help scripting the code for the WGCNA and sPLS analyses. Lastly, we would like to thank Mrs. Ginger Thrash, who connected the Thrash lab to OAR Northwest.

Author Contributions

M.W.H., J.H., F.S., and J.C.T. designed the study, J.H., G.S., P.F., and M.P. collected the data, M.W.H. processed the samples, M.W.H. and J.C.T. analyzed the data, M.W.H. and J.C.T. wrote the manuscript, and all authors contributed to the editing of the manuscript.

Conflict of Interest

The authors declare no competing financial interests.

Literature Cited

1. Richey, J. E., Melack, J. M., Aufdenkampe, A. K., Ballester, V. M. & Hess, L. L. Outgassing from Amazonian rivers and wetlands as a large tropical source of atmospheric CO2. *Nature* **416,** 617–620 (2002).
2. Ensign, S. H. & Doyle, M. W. Nutrient spiraling in streams and river networks. *J. Geophys. Res. Biogeosciences* **111,** (2006).
3. Withers, P. J. A. & Jarvie, H. P. Delivery and cycling of phosphorus in rivers: A review. *Sci. Total Environ.* **400,** 379–395 (2008).

4. Savio, D. *et al.* Bacterial diversity along a 2600 km river continuum. *Environ. Microbiol.* **17,** n/a–n/a (2015).
5. Battin, T. J. *et al.* Biophysical controls on organic carbon fluxes in fluvial networks. *Nat. Geosci.* **2,** 595–595 (2009).
6. Cauwet, G., Hansell, D. A. & Carlson, C. A. in *Biogeochemistry of marine dissolved organic matter.* 579–609 (Academic Press, San Diego, CA, 2002).
7. Cole, J. J. *et al.* Plumbing the global carbon cycle: Integrating inland waters into the terrestrial carbon budget. *Ecosystems* **10,** 171–184 (2007).
8. Cole, J. J. & Caraco, N. F. Carbon in catchments: Connecting terrestrial carbon losses with aquatic metabolism. in *Marine and Freshwater Research* **52,** 101–110 (2001).
9. Dagg, M., Benner, R., Lohrenz, S. & Lawrence, D. Transformation of dissolved and particulate materials on continental shelves influenced by large rivers: Plume processes. *Cont. Shelf Res.* **24,** 833–858 (2004).
10. Turner, R. E. & Rabalais, N. N. Linking landscape and water quality in the Mississippi river basin for 200 years. *Bioscience* **53,** 563–572 (2003).
11. Rabalais, N. N. *et al.* Nutrient Changes in the Mississippi River and System Responses on the Adjacent Continental Shelf. *Estuaries* **19,** 386 (1996).
12. Singh, V. *Application of Frequency and Risk in Water Resources: Proceedings of the International Symposium on Flood Frequency and Risk Analyses, 14–17 May 1986, Louisiana State University, Baton Rouge, USA*. (Springer Science & Business Media, 2012).
13. Dagg, M. J. *et al.* Biogeochemical characteristics of the lower Mississippi River, USA, during June 2003. *Estuaries* **28,** 664–674 (2005).
14. Turner, R. E. & Rabalais, N. N. Suspended sediment, C, N, P, and Si yields from the Mississippi River Basin. *Hydrobiologia* **511,** 79–89 (2004).
15. Schilling, K. E., Chan, K.-S., Liu, H. & Zhang, Y.-K. Quantifying the effect of land use land cover change on increasing discharge in the Upper Mississippi River. *J. Hydrol.* **387,** 343–345 (2010).
16. Staley, C. *et al.* Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River. *Front Microbiol* **5,** 524 (2014).
17. McIsaac, G. F., David, M. B., Gertner, G. Z. & Goolsby, D. a. Nitrate flux in the Mississippi River. *Nature* **414,** 166–167 (2001).
18. Duan, S., Powell, R. T. & Bianchi, T. S. High frequency measurement of nitrate concentration in the Lower Mississippi River, USA. *J. Hydrol.* **519,** 376–386 (2014).
19. Rabalais, N. N. *et al.* Hypoxia in the northern Gulf of Mexico: Does the science support the Plan to Reduce, Mitigate, and Control Hypoxia? *Estuaries and Coasts* **30,** 753–772 (2007).
20. Bianchi, T. S. *et al.* The science of hypoxia in the northern Gulf of Mexico: A review. *Science of the Total Environment* **408,** 1471–1484 (2010).
21. Rabalais, N. N., Turner, R. E. & Wiseman, W. J. Gulf of Mexico Hypoxia, a.K.a. 'the Dead Zone'. *Annu. Rev. Ecol. Syst.* **33,** 235–263 (2002).

765 22. Bristow, L. A. *et al.* Biogeochemical and metagenomic analysis of nitrite
766 accumulation in the Gulf of Mexico hypoxic zone. *Limnol. Oceanogr.* **60,**
767 1733–1750 (2015).
768 23. Lemke, M. J. *et al.* Description of freshwater bacterial assemblages from
769 the upper Paran river floodpulse system, Brazil. *Microb. Ecol.* **57,** 94–103
770 (2009).
771 24. Christian Winter Gerhard Kavka, Robert L. Mach, and, T. H. & Farnleitner,
772 A. H. Longitudinal Changes in the Bacterial Community Composition of the
773 Danube \rRiver: a Whole-River Approach. *AEM* (2007).
774 25. Kolmakova, O. V., Gladyshev, M. I., Rozanov, A. S., Peltek, S. E. &
775 Trusova, M. Y. Spatial biodiversity of bacteria along the largest Arctic river
776 determined by next-generation sequencing. *FEMS Microbiol. Ecol.* **89,**
777 442–450 (2014).
778 26. Staley, C. *et al.* Core functional traits of bacterial communities in the Upper
779 Mississippi River show limited variation in response to land cover. *Front.*
780 *Microbiol.* **5,** 414 (2014).
781 27. Van Rossum, T. *et al.* Year-long metagenomic study of river microbiomes
782 across land use and water quality . *Front. Microbiol.* **6** , 1–15
783 (2015).
784 28. Meziti, A., Tsementzi, D., Ar. Kormas, K., Karayanni, H. & Konstantinidis,
785 K. T. Anthropogenic effects on bacterial diversity and function along a river-
786 to-estuary gradient in Northwest Greece revealed by metagenomics.
787 *Environmental Microbiology* (2016). doi:10.1111/1462-2920.13303
788 29. Read, D. S. *et al.* Catchment-scale biogeography of riverine
789 bacterioplankton. *ISME J* **9,** 516–526 (2015).
790 30. Blanchet, M. *et al.* When riverine dissolved organic matter (DOM) meets
791 labile DOM in coastal waters: changes in bacterial community activity and
792 composition. *Aquat. Sci.* (2016). doi:10.1007/s00027-016-0477-0
793 31. Ruiz-González, C. *et al.* Differences in organic matter and bacterioplankton
794 between sections of the largest Arctic river: Mosaic or continuum?. *Front.*
795 *Microbiol.* **6,** 196–206 (2015).
796 32. Zeglin, L. H. Stream microbial diversity in response to environmental
797 changes: review and synthesis of existing research. *Front. Microbiol.* **6,**
798 454 (2015).
799 33. Newton, R. J. & McLellan, S. L. A unique assemblage of cosmopolitan
800 freshwater bacteria and higher community diversity differentiate an
801 urbanized estuary from oligotrophic Lake Michigan. *Front. Microbiol.* **6,** 1–
802 13 (2015).
803 34. Zwart, G., Crump, B. C., Kamst-van Agterveld, M. P., Hagen, F. & Han, S.
804 K. Typical freshwater bacteria: An analysis of available 16S rRNA gene
805 sequences from plankton of lakes and rivers. *Aquat. Microb. Ecol.* **28,** 141–
806 155 (2002).
807 35. Staley, C. *et al.* Application of Illumina next-generation sequencing to
808 characterize the bacterial community of the Upper Mississippi River. *J.*

Appl. Microbiol. **115,** 1147–1158 (2013).

36. RL, V., GW, M., KW, C., JR, S. & CE, C. River continuum concept. *Can J Fish Aquat Sci* **37,** 130–137 (1980).
37. Fortunato, C. S., Herfort, L., Zuber, P., Baptista, A. M. & Crump, B. C. Spatial variability overwhelms seasonal patterns in bacterioplankton communities across a river to ocean gradient. *ISME J* **6,** 554–563 (2012).
38. Besemer, K. *et al.* Headwaters are critical reservoirs of microbial diversity for fluvial networks. *Proc. Biol. Sci.* **280,** 20131760 (2013).
39. Jackson, C. R., Millar, J. J., Payne, J. T. & Ochs, C. a. Free-living and particle-associated bacterioplankton in large rivers of the Mississippi River Basin demonstrate biogeographic patterns. *Appl. Environ. Microbiol.* **80,** 7186–7195 (2014).
40. Crump, B. C., Armbrust, E. V. & Baross, J. A. Phylogenetic Analysis of Particle-Attached and Free-Living Bacterial Communities in the Columbia River, Its Estuary, and the Adjacent Coastal Ocean. *Appl. Environ. Microbiol.* **65,** 3192–3204 (1999).
41. Riemann, L. & Winding, A. Community dynamics of free-living and particle-associated bacterial assemblages during a freshwater phytoplankton bloom. *Microb. Ecol.* **42,** 274–285 (2001).
42. D'Ambrosio, L., Ziervogel, K., Macgregor, B., Teske, A. & Arnosti, C. Composition and enzymatic function of particle-associated and free-living bacteria: a coastal/offshore comparison. *ISME J.* 1–13 (2014). doi:10.1038/ismej.2014.67
43. Allgaier, M. & Grossart, H.-P. Seasonal dynamics and phylogenetic diversity of free-living and particle-associated bacterial communities in four lakes in northeastern Germany. *Aquat. Microb. Ecol.* **45,** 115–128 (2006).
44. Crump, B. C. & Baross, J. A. Particle-attached bacteria and heterotrophic plankton associated with the Columbia River estuarine turbidity maxima. *Part. Bact. heterotrophic Plankt. Assoc. with Columbia River Estuar. Turbid. maxima* (1996).
45. Grossart, H. P. Ecological consequences of bacterioplankton lifestyles: Changes in concepts are needed. *Environmental Microbiology Reports* **2,** 706–714 (2010).
46. Freimann, R., Bürgmann, H., Findlay, S. E. G. & Robinson, C. T. Hydrologic linkages drive spatial structuring of bacterial assemblages and functioning in alpine floodplains TL - 6. *Front. Microbiol.* **6 VN - re,** (2015).
47. Niño-García, J. P., Ruiz-González, C. & Giorgio, P. A. del. Interactions between hydrology and water chemistry shape bacterioplankton biogeography across boreal freshwater networks. *ISME J.* (2016). doi:10.1038/ismej.2015.226
48. Pierson, S. M., Rosenbaum, B. J., McKay, L. D. & Dewald, T. G. *Strahler Stream Order and Strahler Calculator Values in NHDPlus*. (2008).
49. Gladyshev, M. I. *et al.* Differences in organic matter and bacterioplankton between sections of the largest Arctic river: Mosaic or continuum? *Limnol.*

Oceanogr. **60,** 1314–1331 (2015).

50. Ghai, R. *et al.* Metagenomics of the water column in the pristine upper course of the Amazon river. *PLoS One* **6,** (2011).
51. Staley, C. *et al.* Bacterial community structure is indicative of chemical inputs in the Upper Mississippi River. *Front Microbiol* **5,** 524 (2014).
52. Staley, C. *et al.* Species sorting and seasonal dynamics primarily shape bacterial communities in the Upper Mississippi River. *Sci. Total Environ.* **505,** 435–45 (2015).
53. Crump, B. C., Hopkinson, C. S., Sogin, M. L. & Hobbie, J. E. Microbial Biogeography along an Estuarine Salinity Gradient: Combined Influences of Bacterial Growth and Residence Time. *Appl. Environ. Microbiol.* **70,** 1494–1505 (2004).
54. Russell, T. A. & Weller, L. *State of the River Report: Water Quality and River Health in the Metro Mississippi River*. (Friends of the Mississippi River, 2013).
55. Karr, J. R. Biological integrity: a long-neglected aspect of water resource management. *Ecological Applications* **1,** 66–84 (1991).
56. Sims, A., Zhang, Y., Gajaraj, S., Brown, P. B. & Hu, Z. Toward the development of microbial indicators for wetland assessment. *Water Research* **47,** 1711–1725 (2013).
57. Karr, J. R. Assessment of biotic integrity using fish communities. *Fisheries* **6,** 21–27 (1981).
58. Thrash, J. C. & Coates, J. D. in *Bergey's Manual® of Systematic Bacteriology* 725–735 (Springer, 2010).
59. Kielak, A. M., Barreto, C. C., Kowalchuk, G. A., van Veen, J. A. & Kuramae, E. E. The Ecology of Acidobacteria: Moving beyond Genes and Genomes. *Front. Microbiol.* **7,** 744 (2016).
60. Anderson, I. *et al.* Genome sequence of the homoacetogenic bacterium Holophaga foetida type strain (TMBS4 T). *Stand. Genomic Sci.* **6,** 174 (2012).
61. Allen, M. B. & Arnon, D. I. Studies on nitrogen-fixing blue-green algae. I. Growth and nitrogen fixation by Anabaena cylindrica Lemm. *Plant Physiol.* **30,** 366 (1955).
62. Van Geel, B., Mur, L. R., Ralska-Jasiewiczowa, M. & Goslar, T. Fossil akinetes of Aphanizomenon and Anabaena as indicators for medieval phosphate-eutrophication of Lake Gosciaz (Central Poland). *Rev. Palaeobot. Palynol.* **83,** 97–105 (1994).
63. Wood, S. A., Prentice, M. J., Smith, K. & Hamilton, D. P. Low dissolved inorganic nitrogen and increased heterocyte frequency: precursors to Anabaena planktonica blooms in a temperate, eutrophic reservoir. *J. Plankton Res.* **32,** 1315–1325 (2010).
64. Addison, S. L., Foote, S. M., Reid, N. M. & Lloyd-Jones, G. Novosphingobium nitrogenifigens sp. nov., a polyhydroxyalkanoate-accumulating diazotroph isolated from a New Zealand pulp and paper

wastewater. *Int. J. Syst. Evol. Microbiol.* **57,** 2467–2471 (2007).
65. Li, H.-F. *et al.* Novosphingobium sediminis sp. nov., isolated from the sediment of a eutrophic lake. *J. Gen. Appl. Microbiol.* **58,** 357–362 (2012).
66. Trusova, M. Y. & Gladyshev, M. I. Phylogenetic diversity of winter bacterioplankton of eutrophic siberian reservoirs as revealed by 16S rRNA gene sequence. *Microb. Ecol.* **44,** 252–259 (2002).
67. Cebron, A. & Garnier, J. Nitrobacter and Nitrospira genera as representatives of nitrite-oxidizing bacteria: detection, quantification and growth along the lower Seine River (France). *Water Res.* **39,** 4979–4992 (2005).
68. Korajkic, A. *et al.* Changes in bacterial and eukaryotic communities during sewage decomposition in Mississippi river water TL - 69. *Water Res.* **69 VN - r,** 30–39 (2015).
69. Necchi Jr, O. in *River Algae* 153–158 (Springer, 2016).
70. Jansson, M., Blomqvist, P., Jonsson, A. & BergstrÖm, A. Nutrient limitation of bacterioplankton, autotrophic and mixotrophic phytoplankton, and heterotrophic nanoflagellates in Lake Örträsket. *Limnol. Oceanogr.* **41,** 1552–1559 (1996).
71. Caron, D. A., Porter, K. G. & Sanders, R. W. Carbon, nitrogen, and phosphorus budgets for the mixotrophic phytoflagellate Poterioochromonas malhamensis (Chrysophyceae) during bacterial ingestion. *Limnol. Oceanogr.* **35,** 433–443 (1990).
72. Smith, M. W. *et al.* Seasonal Changes in Bacterial and Archaeal Gene Expression Patterns across Salinity Gradients in the Columbia River Coastal Margin. *PLoS One* **5,** e13312 (2010).
73. Fortunato, C. S. *et al.* Determining indicator taxa across spatial and seasonal gradients in the Columbia River coastal margin. *ISME J* **7,** 1899–1911 (2013).
74. Vorosmarty, C. J. *et al.* Global threats to human water security and river biodiversity. *Nature* **467,** 555–561 (2010).
75. Weiss, S. *et al.* Tracking down the sources of experimental contamination in microbiome studies. *Genome Biol.* **15,** 564 (2014).
76. Salter, S. J. *et al.* Reagent and laboratory contamination can critically impact sequence-based microbiome analyses. *BMC Biol.* **12,** 87 (2014).
77. Thrash, J. C., Weckhorst, J. L. & Pitre, D. M. in *Protocols for Metagenomic Library Generation and Analysis in Petroleum Hydrocarbon Microbe Systems* 1–22 (Humana Press, 2015). doi:10.1007/8623_2015_67
78. Caporaso, J. G. *et al.* Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J.* **6,** 1621–1624 (2012).
79. Amaral-Zettler, L. A., McCliment, E. A., Ducklow, H. W. & Huse, S. M. A Method for Studying Protistan Diversity Using Massively Parallel Sequencing of V9 Hypervariable Regions of Small-Subunit Ribosomal RNA Genes. *PLoS One* **4,** e6372 (2009).
80. Schloss, P. D. *et al.* Introducing mothur: Open-source, platform-

independent, community-supported software for describing and comparing microbial communities. *Appl. Environ. Microbiol.* **75,** 7537–7541 (2009).
81. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: Improved data processing and web-based tools. *Nucleic Acids Res.* **41,** 590–596 (2013).
82. Pruesse, E. *et al.* SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res.* **35,** 7188–7196 (2007).
83. Edgar, R. C., Haas, B. J., Clemente, J. C., Quince, C. & Knight, R. UCHIME improves sensitivity and speed of chimera detection. *Bioinformatics* **27,** 2194–2200 (2011).
84. Oksanen, J. *et al.* vegan: Community Ecology Package. R package version 2.2-1. *R package version* **1,** R package version 2.2–1. (2015).
85. R, C. T. *R: A language and environment for statistical computing* (2013).
86. McMurdie, P. J. & Holmes, S. phyloseq: An R Package for Reproducible Interactive Analysis and Graphics of Microbiome Census Data. *PLoS One* **8,** e61217 (2013).
87. Love, M. I., Anders, S. & Huber, W. Differential analysis of count data - the DESeq2 package. *Genome Biol.* **15,** 550 (2014).
88. Learman, D. R. *et al.* Biogeochemical and microbial variation across 5500 km of Antarctic surface sediment implicates organic matter as a driver of benthic community structure. *Front. Microbiol.* **7,** (2016).
89. Langfelder, P. & Horvath, S. WGCNA: an R package for weighted gene co-expression network analysis. *BMC Bioinformatics* **9,** 559 (2008).
90. Guidi, L. *et al.* Plankton networks driving carbon export in the oligotrophic ocean. *Nature* in review (2015). doi:10.1038/nature16942
91. Chong, I. G. & Jun, C. H. Performance of some variable selection methods when multicollinearity is present. *Chemom. Intell. Lab. Syst.* **78,** 103–112 (2005).
92. Mevik, B.-H. & Wehrens, R. The pls Package: Principle Component and Partial Least Squares Regression in R. *J. Stat. Softw.* **18,** 1–24 (2007).

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. C) An OAR Northwest rowboat dwarfed by a paddleboat.

Figure 2. Non-metric multidimensional scaling (NMDS) results for whole community correlations with 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. 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 5. 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 6. PLS results for the 0.2-2.7 μm 16S rRNA gene community for selected submodules with nitrate and phosphate and a VIP score > 1. Correlation of submodule OTUs to nitrate (A) and phosphate (B), according to the number of co-correlations (node centrality). 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.

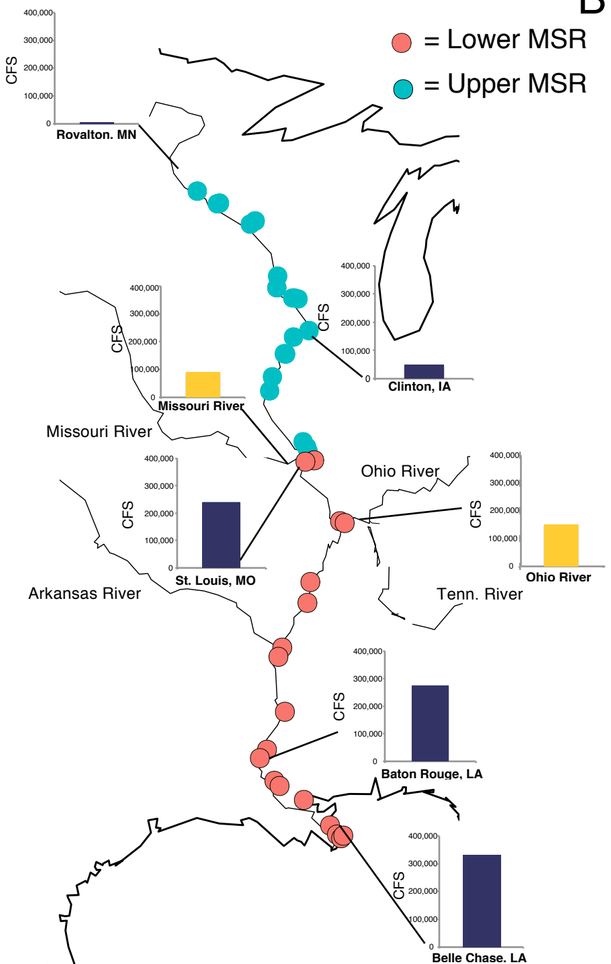
Figure 7. PLS results for the > 2.7 μm 16S rRNA gene community for selected submodules with nitrate and phosphate and a VIP score of > 1. Correlation of submodule OTUs to nitrate (A) and phosphate (B), according to the number of co-correlations (node centrality). Circle size is proportional to VIP scores, 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 8. PLS results for the 0.2-2.7 μm 16S rRNA gene community for selected submodules with nitrate and phosphate and a VIP score of > 1. Correlation of submodule OTUs to nitrate (A) and phosphate (B), according to the number of co-correlations (node centrality). 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.

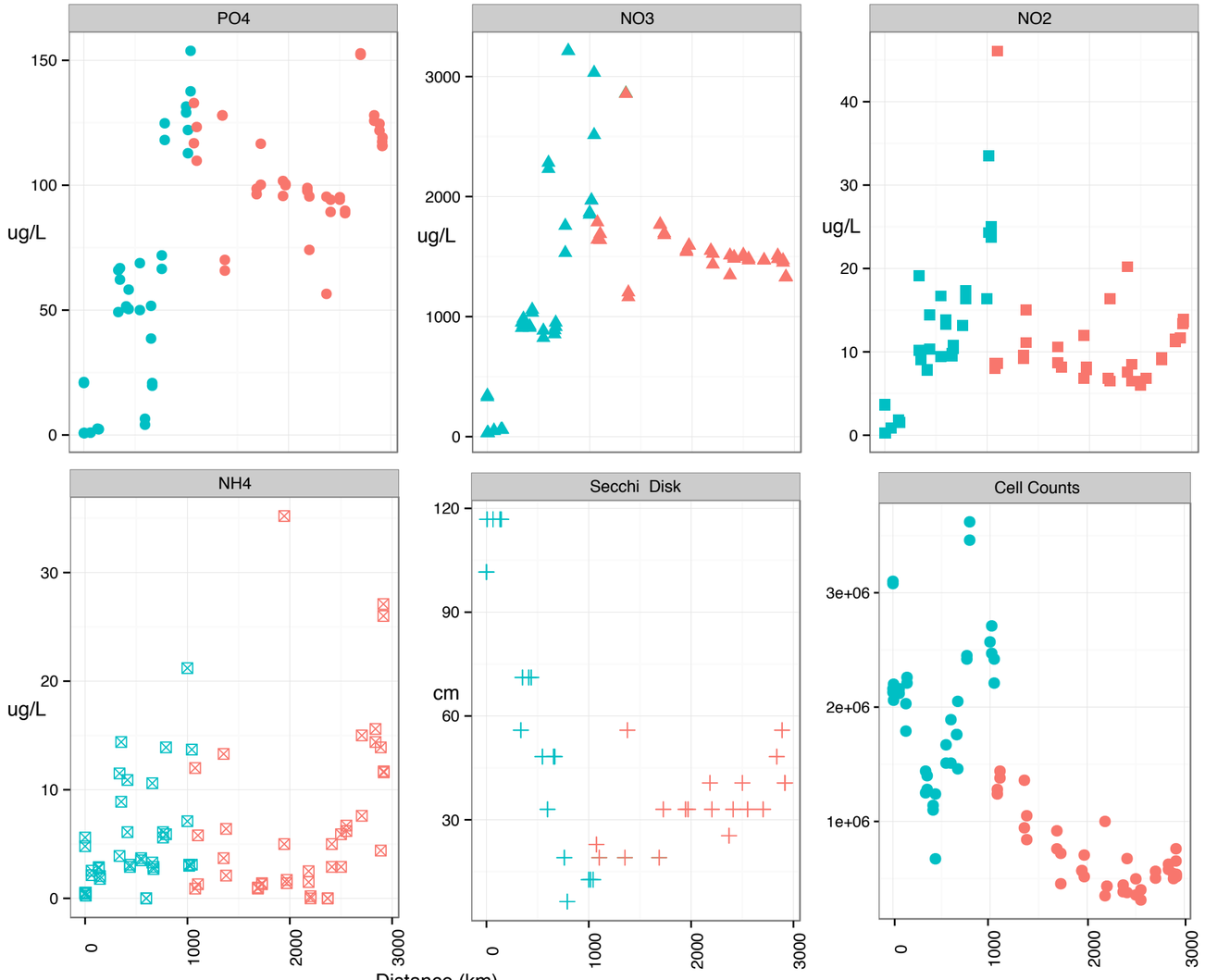
Figure 9. PLS results for the > 2.7 μm 18S rRNA gene community for selected submodules with nitrate and phosphate and a VIP score of > 1. Correlation of submodule OTUs to nitrate (A) and phosphate (B), according to the number of co-correlations (node centrality). Circle size is proportional to

1029 VIP scores, with top 10 VIP scoring and top node centrality OTUs labeled
1030 with their highest-resolution taxonomic classification and OTU number.
1031 Colors represent the taxonomic classification the phylum level.

A

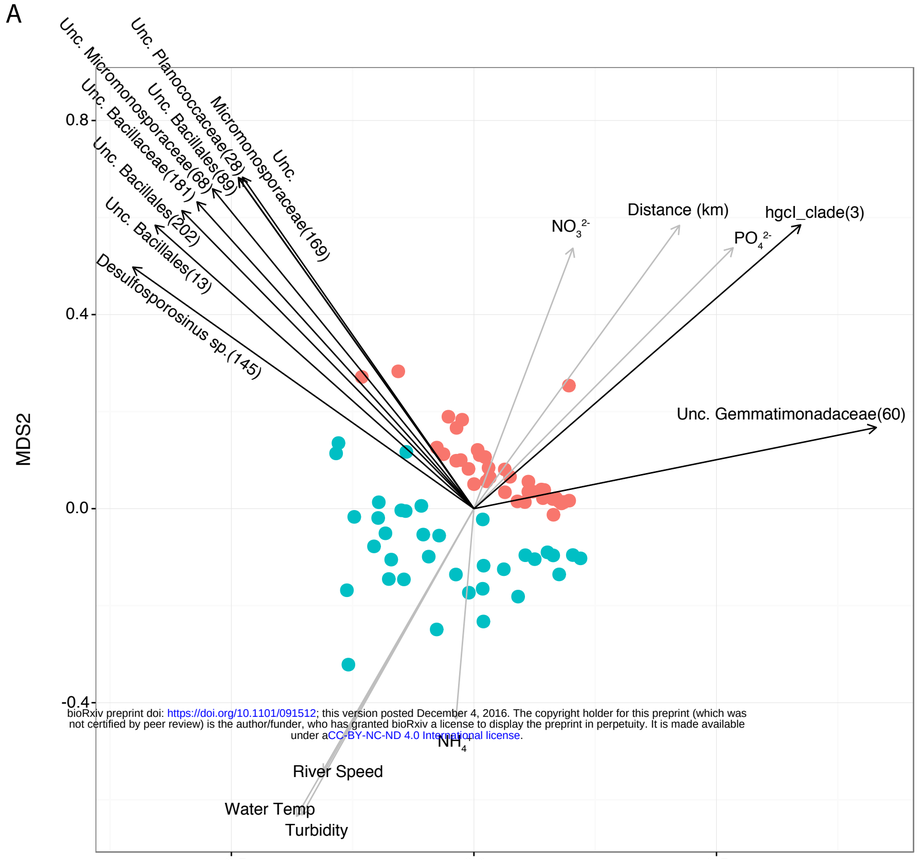

= Lower MSR
= Upper MSR
CFS
Rovalton. MN
Missouri River
Clinton, IA
Missouri River
Ohio River
St. Louis, MO
Ohio River
Arkansas River
Tenn. River
Baton Rouge, LA
Belle Chase. LA

B

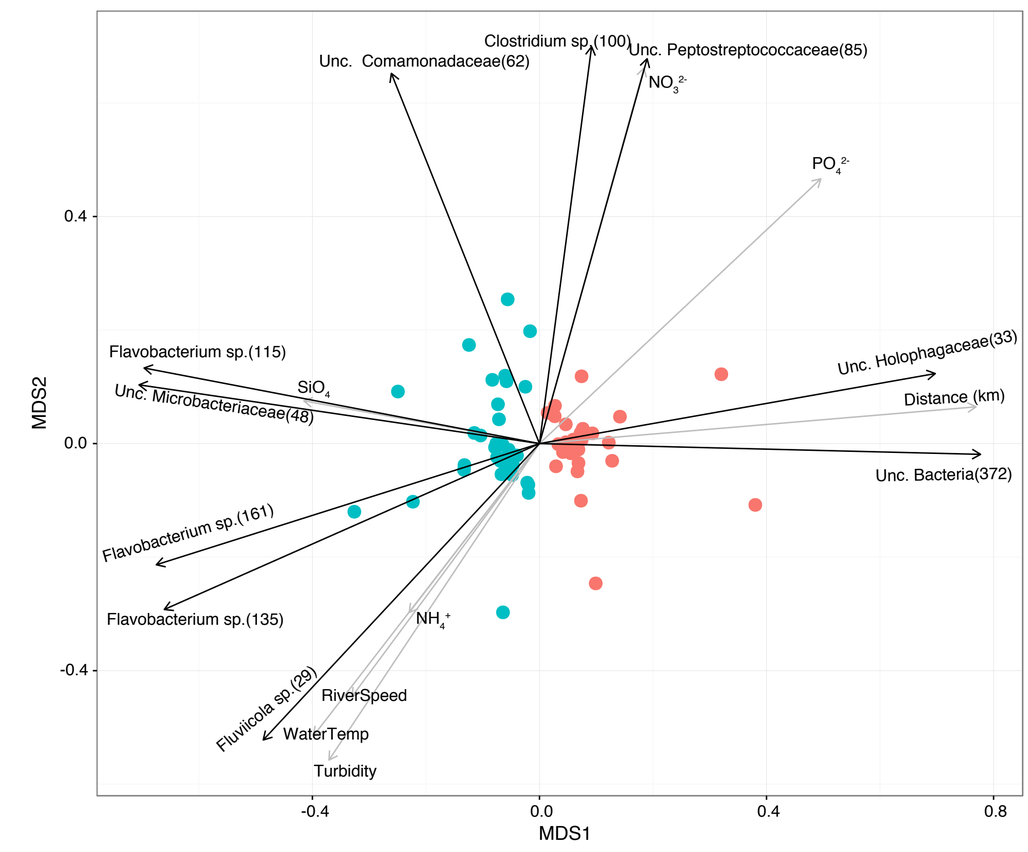

PO4
NO3
NO2
NH4
Secchi Disk
Cell Counts
ug/L
cm
Distance (km)

C

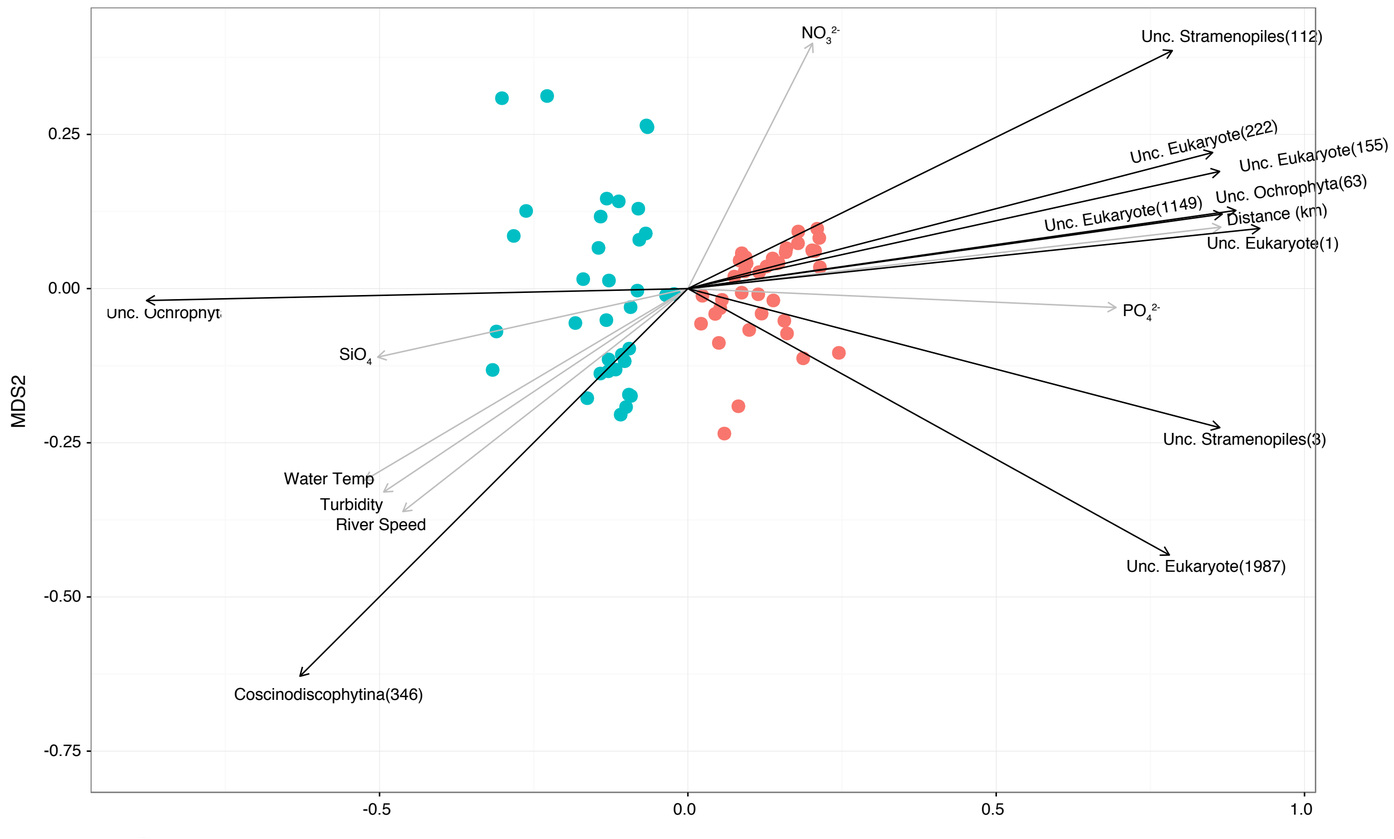
A



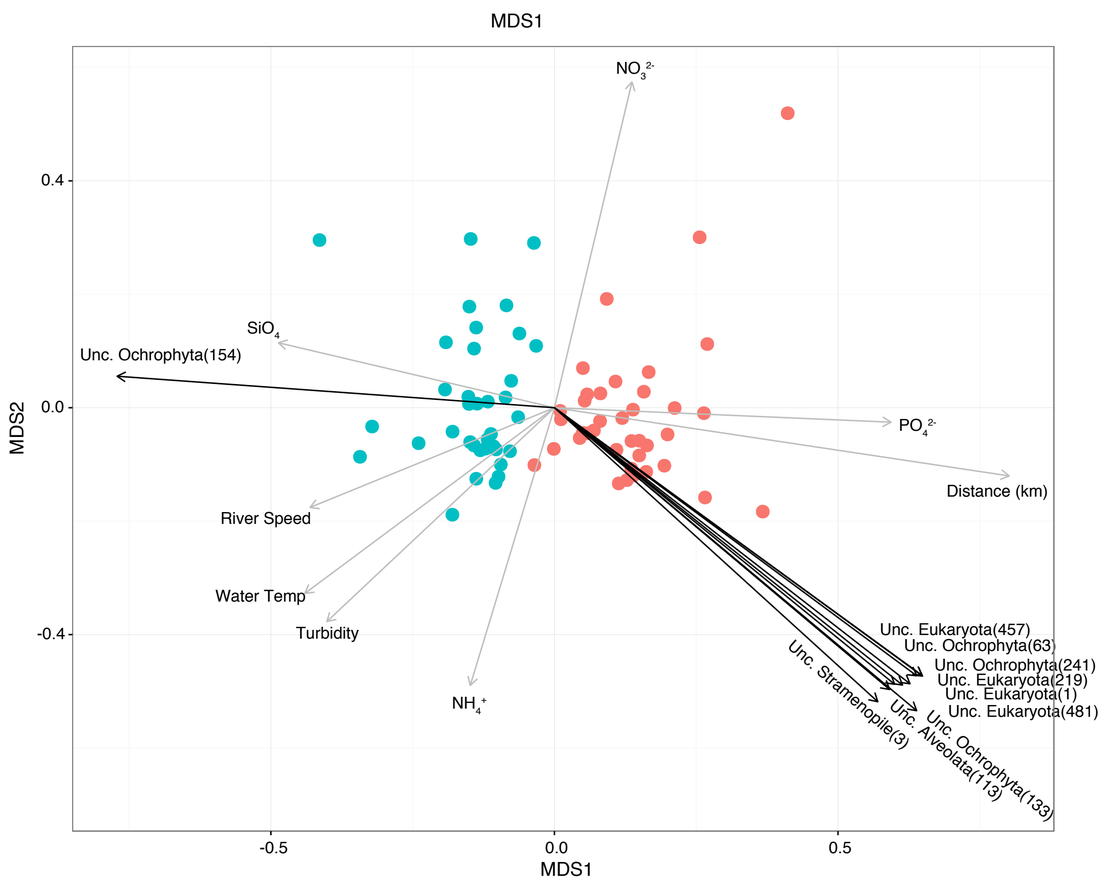
B



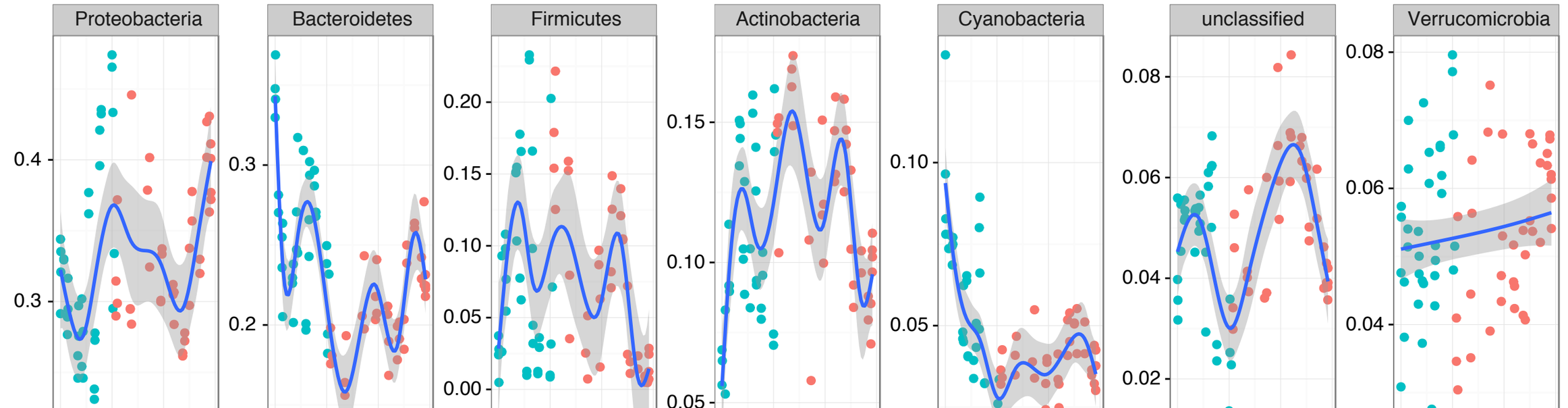
C



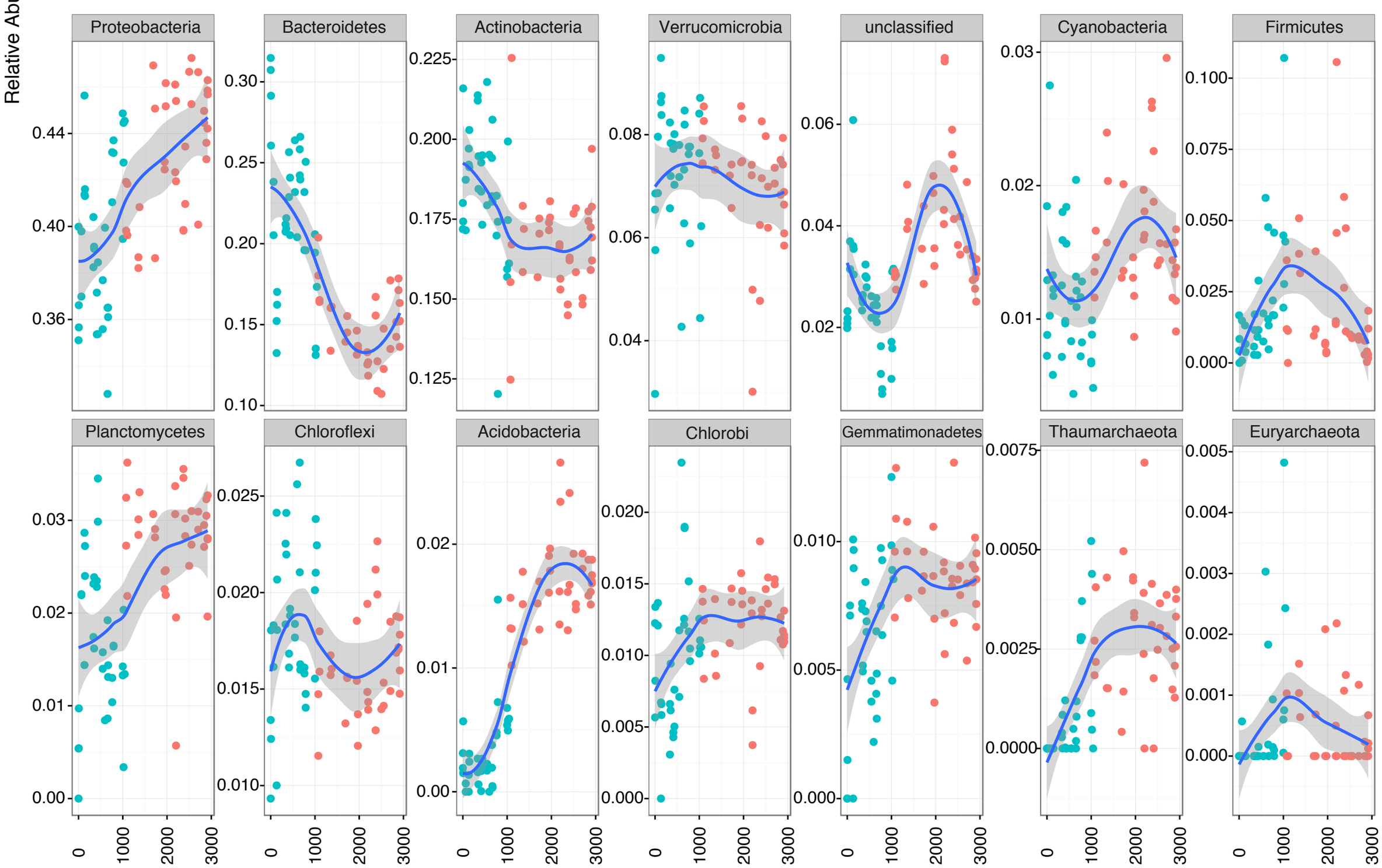
D

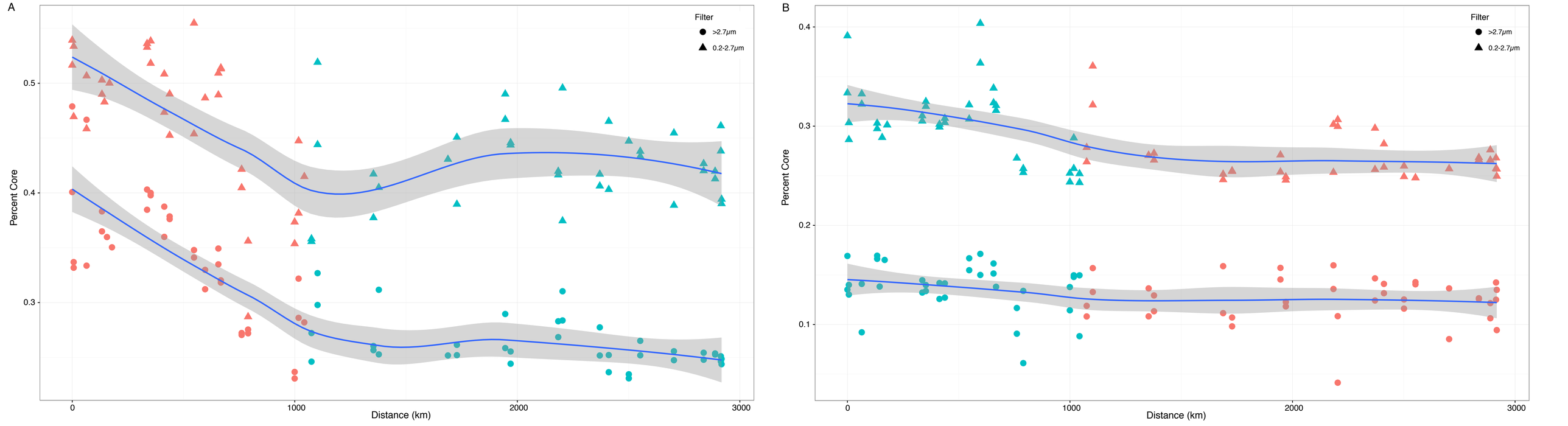


A

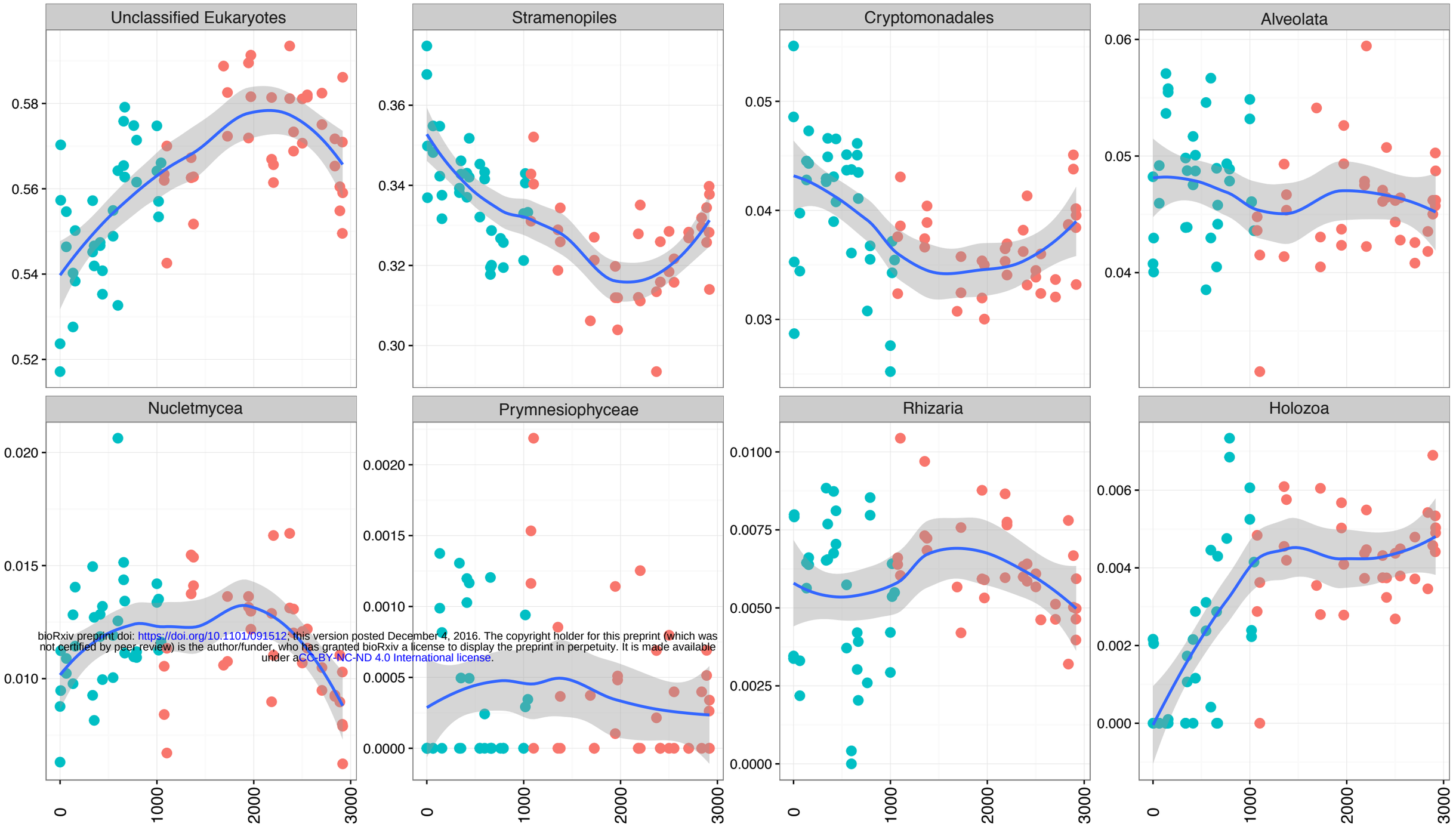


B

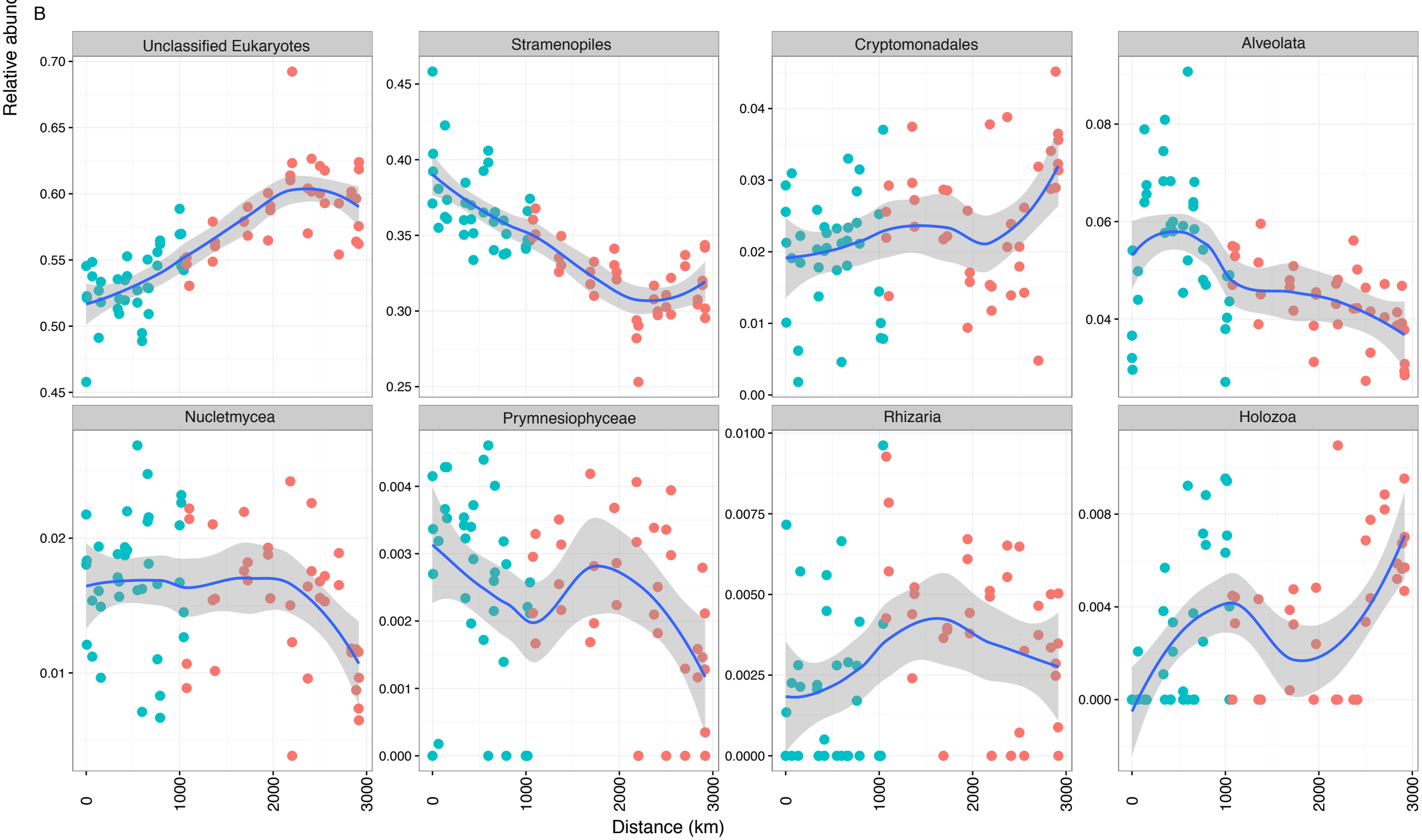



A
Filter
>2.7μm
0.2-2.7μm
Percent Core
0.5
0.4
0.3
0
1000
2000
3000
Distance (km)
B
Filter
>2.7μm
0.2-2.7μm
Percent Core
0.4
0.3
0.2
0.1
0
1000
2000
3000
Distance (km)

A

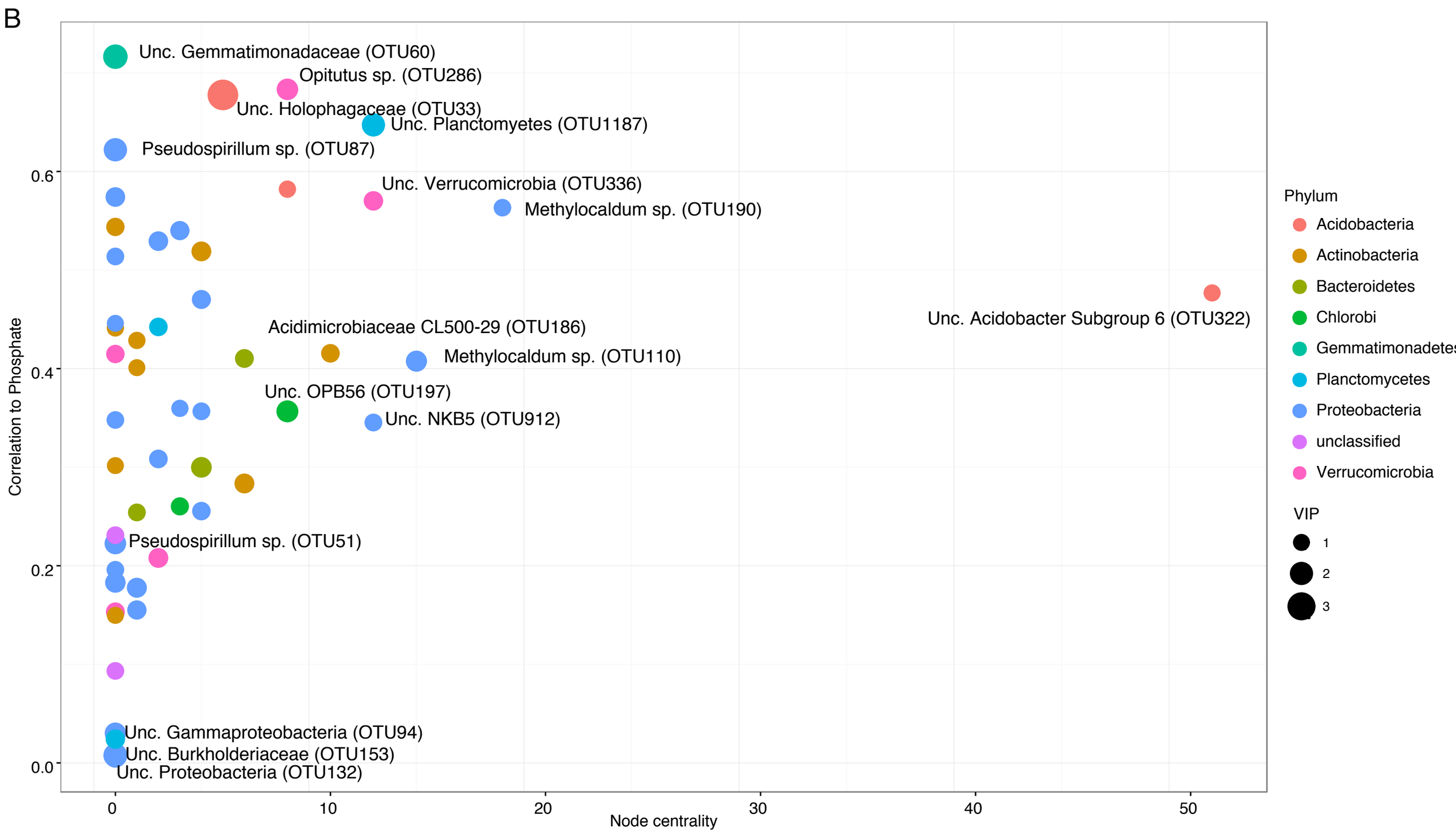


B



A

B



A

Novosphingobium sp. (OTU117)
Zymomonas sp. (OTU164)
Novosphingobium sp. (OTU22)
Altererythrobacter sp. (OTU459)
Mucilaginibacter sp. (OTU129)
Lacibacter sp. (OTU583)
Sphingomonas sp. (OTU111)
Novosphingobium sp. (OTU345)
Unc. Sphingobacteriacea (OTU81)
Unc. Comamonadaceae (OTU62)
Leadbetterella sp. (OTU307)
Flavobacterium sp. (OTU20)
Flavobacterium sp. (OTU54)
Flavobacterium sp. (OTU65)
Undibacterium sp. (OTU188)
Anabaena sp. (OTU40)
NS11-12 Marine Group (OTU57)
Unc. Proteobacteria (OTU412)
Unc. Bacteria (OTU270)
Snowella sp. (OTU500)

Correlation to Nitrate
0.4
0.0
-0.4
0
20
40
60
Node centrality

Phylum
Actinobacteria
Bacteroidetes
Cyanobacteria
Proteobacteria
Unc. Bacteria
Verrucomicrobia

VIP
1.0
1.5
2.0

B

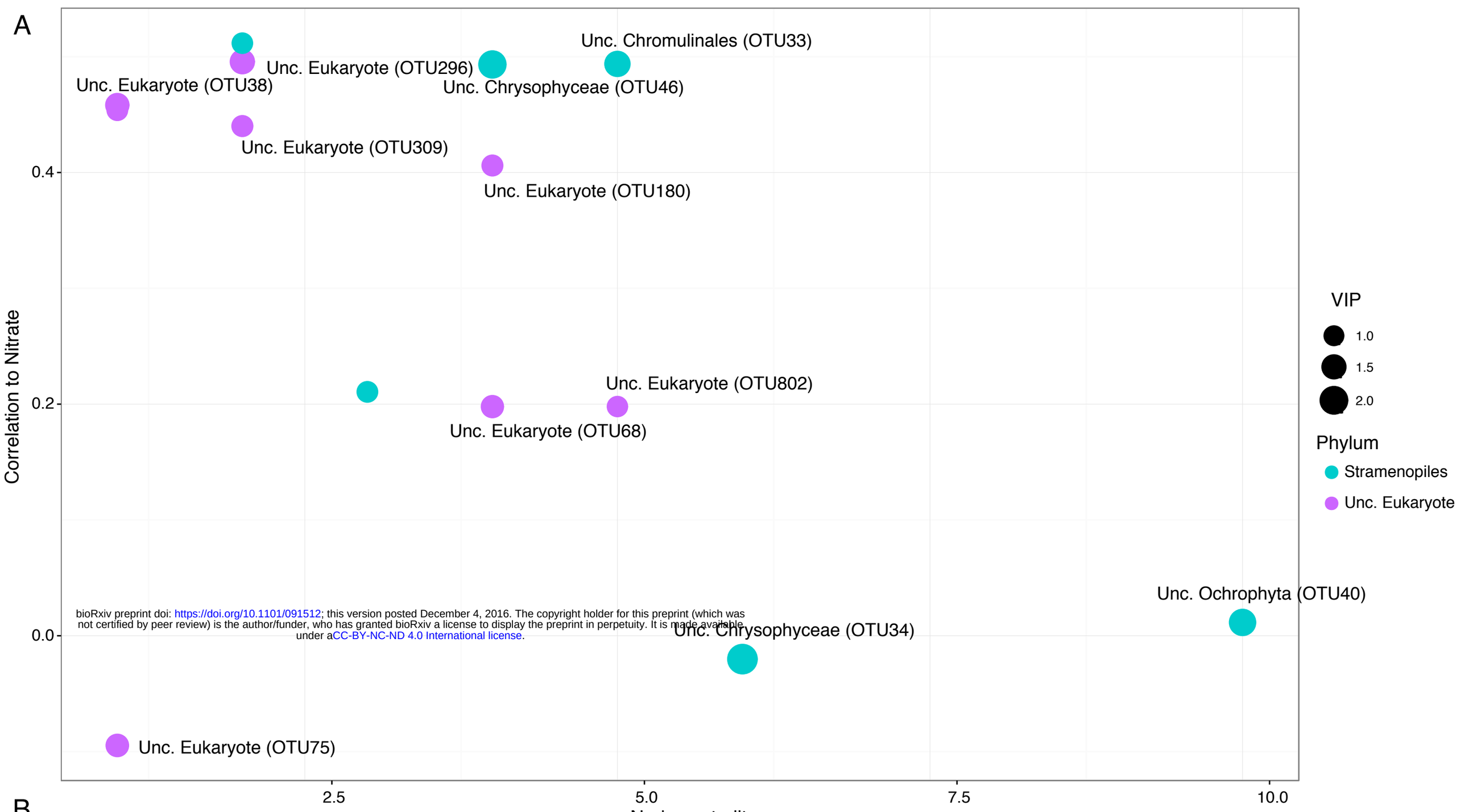
Nitrospira sp. (OTU96)
Unc. Cytophagacea (OTU222)
Fluviicola sp. (OTU302)
Tolumonas sp. (OTU252)
Unc. Comamonadaceae (OTU97)
Woodsholea sp. (OTU385)
Arcicella sp. (OTU258)
Unc. Gammaproteobacteria (OTU170)
Halioglobus sp. (OTU573)
Unc. Rhodobacteraceae (OTU121)
Shewenella sp. (OTU362)

Correlation to Phosphate
0.6
0.5
0.4
0.3
0.2
0
10
20
30
Node centrality

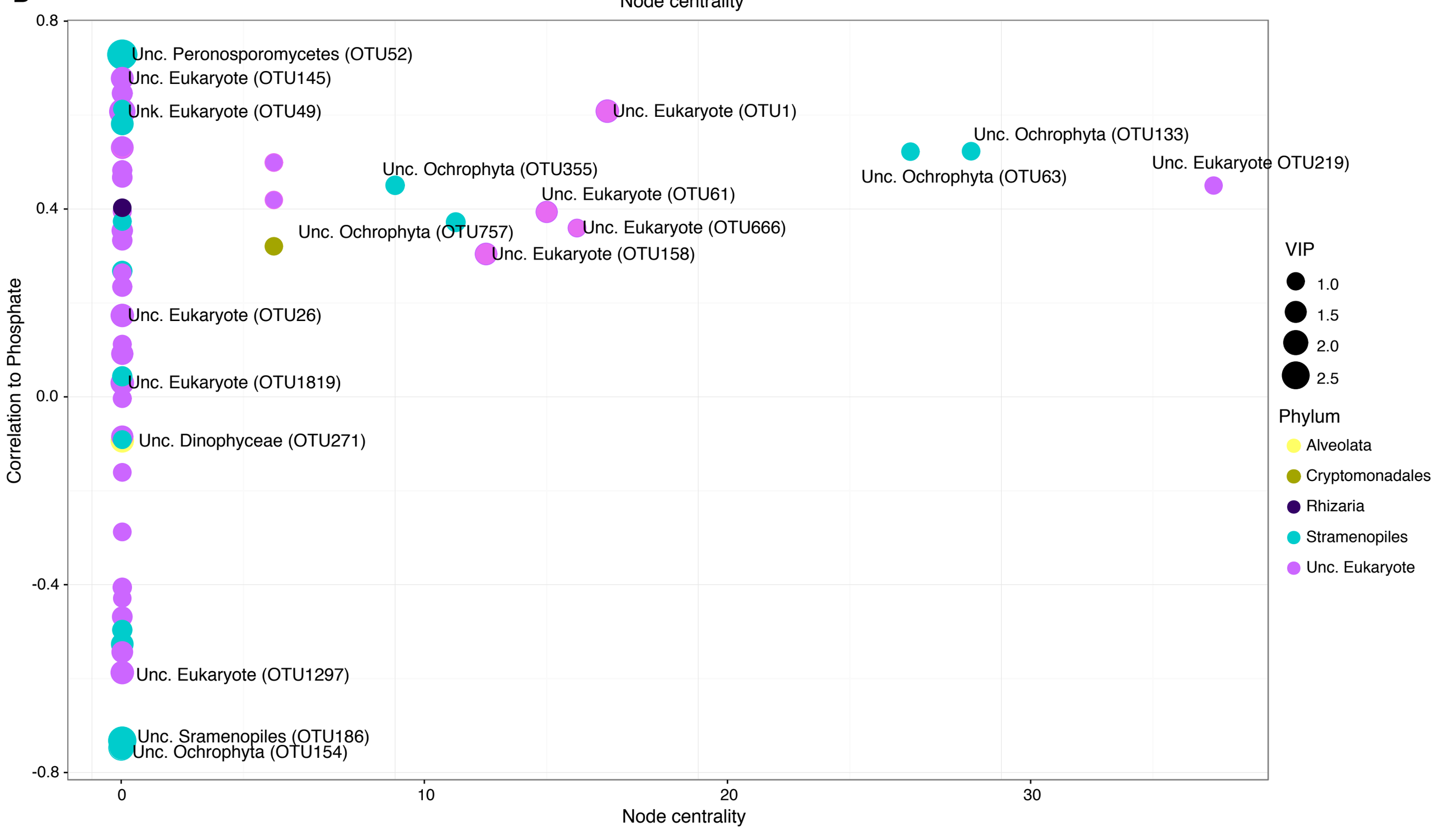
VIP
1.00
1.25
1.50
1.75

Phylum
Acidobacteria
Bacteroidetes
Chloroflexi
Nitrospirae
Proteobacteria
Verrucomicrobia

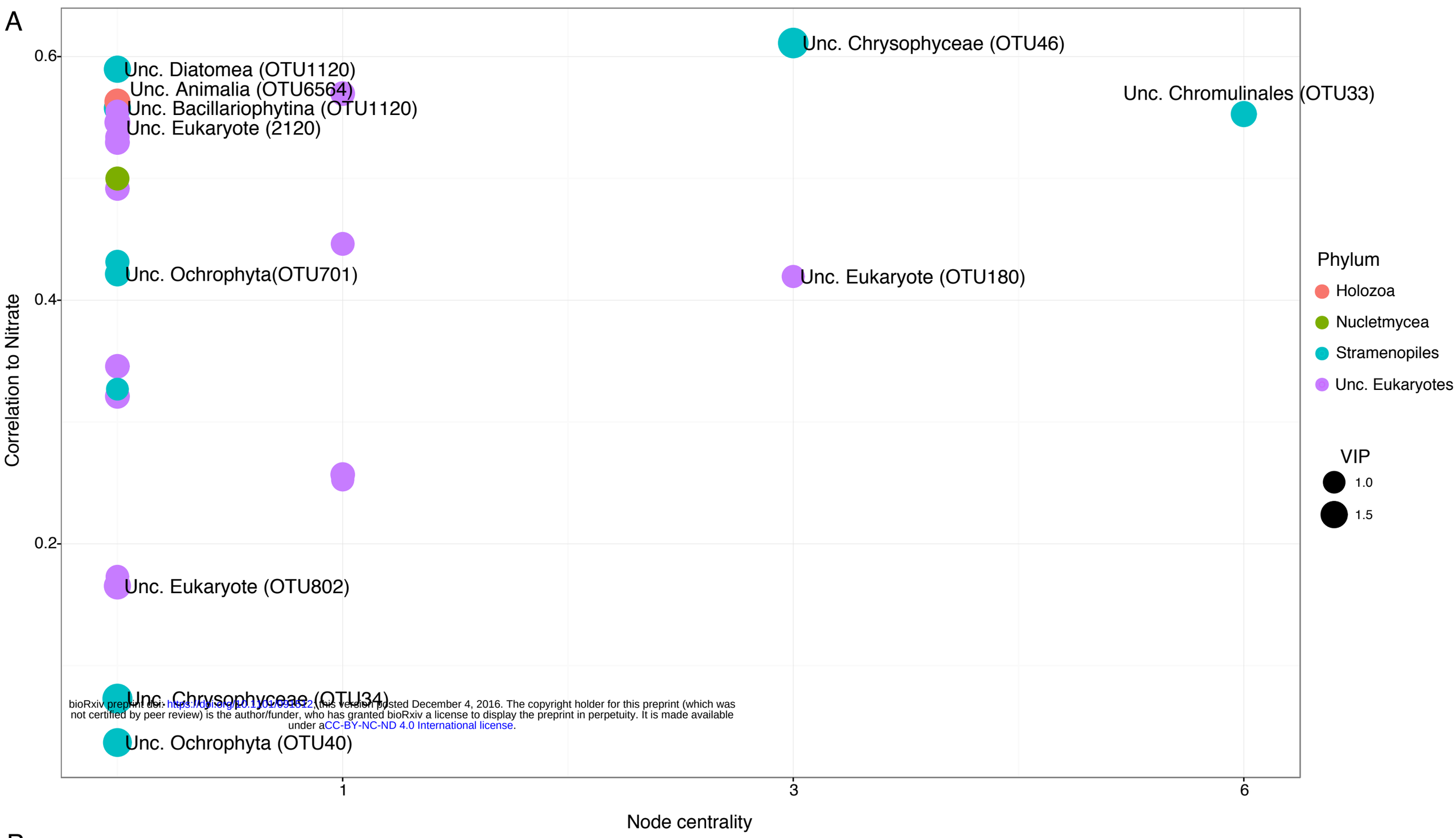
A



B



A



B