Environmental DNA reveals the structure of phytoplankton assemblages along a 2900-km transect in the Mississippi River

Authors

Joseph M. Craine[1,*], Michael W. Henson[2], J. Cameron Thrash[2], Jordan Hanssen[3], Greg Spooner[3], Patrick Fleming[3], Markus Pukonen[3], Frederick Stahr[4], Sarah Spaulding[5], Noah Fierer[1,6,7]

Affiliations

1. Jonah Ventures, 1600 Range St. #201, Boulder, CO, 80301 USA.

2. Department of Biological Sciences, Louisiana State University, Baton Rouge, LA, USA.

3. O.A.R. Northwest, Seattle, WA 98103, U.S.A.

4. School of Oceanography, University of Washington, Seattle, WA 98195, USA.

5. Institute of Arctic and Alpine Research, University of Colorado, Boulder, 80309 USA

6. Department of Ecology and Evolutionary Biology, University of Colorado, Boulder, CO 80309, USA

7. Cooperative Institute for Research in Environmental Science, University of Colorado, Boulder, CO 80309

20 *Contact

21 josephmcraine@gmail.com

Abstract

The environmental health of aquatic ecosystems is critical to society, yet traditional assessments of water quality have limited utility for some bodies of water such as large rivers. Sequencing of environmental DNA (eDNA) has the potential to complement if not replace traditional sampling of biotic assemblages for the purposes of reconstructing aquatic assemblages and, by proxy, assessing water quality. Despite this potential, there has been little testing of the ability of eDNA to reconstruct assemblages and their absolute and relative utility to infer water quality metrics. Here, we reconstruct phytoplankton communities by amplifying and sequencing DNA from a portion of the 23S rRNA region from filtered water samples along a 2900-km transect in the Mississippi River. Across the entire length, diatoms dominated the assemblage (72.6%) followed by cryptophytes (8.7%) and cyanobacteria (7.0%). There were no general trends in the abundances of these major taxa along the length of the river, but individual taxon abundance peaked in different regions. For example, the abundance of taxa genetically similar to *Melosira tropica* peaked at approximately 60% of all reads 2750 km upstream from the Gulf of Mexico, while taxa similar to *Skeletonema marinoi* began to increase below the confluence with the Missouri River until it reached approximately 30% of the reads at the Gulf of Mexico. There were four main clusters of samples based on phytoplankton abundance, two above the confluence with the Missouri and two below. Phytoplankton abundance was a poor predictor of NH_4^+ concentrations in the water, but predicted 61% and 80% of the variation in observed NO_3^- and PO_4^{3-} concentrations, respectively. Phytoplankton richness increased with increasing distance along the river, but was best explained by phosphate concentrations and water clarity. Along the Mississippi transect, there was similar structure to phytoplankton and bacterial assemblages, indicating that the two sets of organisms are responding to similar environmental

45 factors. In all, the research here demonstrates the potential utility of metabarcoding for

46 reconstructing aquatic assemblages, which might aid in conducting water quality assessments.

47

Introduction

Inland freshwater systems provide vital services of drinking water, habitat for fisheries, irrigation for agriculture and recreation (Davies and Jackson 2006, American Sportfishing Association 2015). Yet, the ecological status of lakes, rivers, streams, and reservoirs is increasingly threatened by agriculture, roads, industry, mining, human waste, urbanization, and deforestation (Malmqvist and Rundle 2002, US Environmental Protection Agency 2015) Effective monitoring of water quality and the causes of water quality impairment are critical steps to maintaining freshwater resources, preventing further degradation, and guiding restoration efforts. Quantifying the state and dynamics of aquatic ecosystems is often best done indirectly by quantifying the structure of aquatic assemblages (Palaniappan et al. 2010, Young and Loomis 2014). Because each organism has a unique set of ecological traits and responds uniquely to environmental conditions, their abundance in waters is an indicator of environmental conditions such as salinity, temperature, oxygen levels, nutrient supplies, and turbidity (Karr 1999, Schoolmaster et al. 2012).

Fish and aquatic invertebrates are the two of the most common indicators quantified for the purpose of inferring water quality (Barbour et al. 1999, Stein et al. 2014a). Yet, assessments of these assemblages are currently labor intensive, slow, expensive, and often imprecise. For example, manually sampling fish or aquatic invertebrate communities can cost approximately US$2500 for a single site, limiting the number of sites that can be sampled (Stein et al. 2014a). Biotic assessments also are less effective for certain types of aquatic ecosystems. For example, even without fiscal constraints, assessments of fish and aquatic insect assemblages of large rivers can be exceedingly difficult. The efficiency of sampling fish in large rivers with traditional electrofishing is low (and seasonally variable) due to a number of factors such as turbidity

71 (Goffaux et al. 2005, Reyjol et al. 2005, Lyon et al. 2014). Standard techniques for kicknetting
72 insects or collecting exuviae do not work on large rivers (Buss et al. 2015).
73
74 One response to the constraints on sampling fish and insects for large rivers is to rely on other
75 organisms such as diatoms for water quality assessments (Kelly and Whitton 1995, Stein et al.
76 2014a). Yet, traditional techniques for visually assessing the relative abundance of taxa such as
77 diatoms is still expensive, subject to taxonomic bias, and constrained by low taxonomic
78 resolution (Zimmermann et al. 2015). Given these constraints, next generation sequencing of
79 environmental DNA has the potential to quantify assemblages of not only fish and aquatic
80 insects, but also smaller organisms such as phytoplankton or bacteria (Mächler et al. 2014, Stein
81 et al. 2014b, Barnes and Turner 2015, Thomsen and Willerslev 2015). To accomplish this, DNA
82 present in the water is filtered and then regions in the genome are amplified and sequenced,
83 providing information on the presence, if not relative abundance, of organisms. Depending on
84 the regions of the genome amplified, different taxonomic groups can be sequenced, including
85 bacteria, phytoplankton, arthropods, fish, and mammals (Jackson et al. 2014, Stein et al. 2014b,
86 Cannon et al. 2016, Deiner et al. 2016, Olds et al. 2016). This potential is coupled with the
87 ability to provide data for less cost, or improved taxonomic specificity, and at a faster rate. For
88 example, water can be filtered and analyzed for environmental DNA at less than a tenth of the
89 cost of traditional biotic assessments.
90
91 Despite this potential, there are few examples of successful application of metabarcoding for
92 reconstructing phytoplankton assemblages and we have not started in earnest to assess whether
93 these reconstructions have value on their own, no less relative to reconstructions generated with

organisms (Hamsher et al. 2013). To better understand the potential of environmental DNA to reconstruct phytoplankton assemblages in a large river, we amplified and sequenced DNA using a 23S rRNA gene region primer pair (Sherwood and Presting 2007) specific to phytoplankton (hereafter, 23S) for 39 sites along over 2900 km of the Mississippi River in addition to its headwaters at Lake Itasca. The Mississippi River is one of the Great Rivers of the US and has been the subject of a number of studies attempting to assess the ecological health of its waters with biological assessment (Angradi et al. 2009, Kireta et al. 2012b, Bellinger et al. 2013). With these data, we examined the patterns of phytoplankton assemblages along the length of the river to determine how the structure and richness of these assemblages changed along the length of the river. As a first test, we compared relationships between the abundance of 23S OTUs and nutrient concentrations in the water. Next, to assess whether the factors structuring phytoplankton were similar to those structuring bacterial assemblages, we compared assemblage structure of 23S and bacterial 16S rRNA gene (hereafter, 16S) OTUs from a previous set of analyses (Henson et al. in review). This was followed with a comparison of the explanatory power of 23S and 16S OTUs to predict nutrient concentrations.

Methods

Sample acquisition

Duplicate water samples were collected from 39 sites along the Mississippi River from September 18, 2014 to November 26, 2014 (Henson et al. in review). The core of these sites spanned 2917 km, from Minneapolis, MN to the Gulf of Mexico. An additional sample was acquired from Lake Itasca, the headwaters of the river. At each site, 120 mL of water was filtered through a 2.7 μm GF/D filter (Whatman GE, New Jersey, USA) and then a 0.2 μm Sterivex filter

(EMD Millipore, Darmstadt, Germany) with a sterile 60 mL syringe (BD, New Jersey, USA). The first 60 mL of flow-through water was collected and saved in an autoclaved, acid-washed 60 mL polycarbonate bottle. Filters and filtrate were stored on ice until they could be shipped to the laboratory for analyses. At each site, light penetration was assessed with a secchi disk (Wildco, Yulee, FL).

Sample processing

DNA was extracted from filters with a MoBio PowerWater DNA kit (MoBio Laboratories, Carlsbad, CA) following the manufacturer's protocol. If there was sufficient DNA remaining from previous analyses (Henson et al. in review), DNA from the two fractions for a site were combined. For some sites, DNA from the two fractions taken from the two replicate samples were combined. Phytoplankton sequences were amplified at the 23S rRNA gene region, which is located on the chloroplast and can amplify DNA from taxa such as cyanobacteria, green algae, and diatoms (Sherwood and Presting, 2007). Initial PCR amplification included Promega Mastermix, forward and reverse primers, gDNA, and DNase/RNase-free H_2O. After an initial 3-minute period at 94°C, DNA was PCR amplified for 40 cycles at 94°C (30 seconds), 55°C (45 seconds) and 72°C (60 seconds), followed by 10 minutes at 72°C. Products were then visualized on an 2% agarose gel. 20µl of the PCR amplicon was used for PCR clean-up using ExoI/SAP reaction. In order to index the amplicons with a unique identifier sequence, the first PCR stage was followed by an indexing 8-cycle PCR reaction to attach 10-bp error-correcting barcodes unique to each sample to the pooled amplicons. These products were again visualized on a 2% agarose gel and checked for band intensity and that amplicons are the correct size. PCR products were purified and normalized using the Life Technologies SequalPrep Normalization kit and

138 samples pooled together. Amplicons were sequenced on an Illumina MiSeq at the University of
139 Colorado Boulder BioFrontiers Sequencing Center running the v2 500-cycle kit.

140 For nutrient analyses, filtrate was previously analyzed colorimetrically for [NH_4^+], [NO_3^-], and
141 [PO_4^{3-}] at the University of Washington Marine Chemistry Laboratory as described in Henson et
142 al. (in review).

143 **Bioinformatic processing**

144 Sequences were demulitplexed using a python script. Paired end reads were then merged using
145 fastq_merge pairs. Since merged reads often extended beyond the amplicon region of the
146 sequencing construct, we used fastx_clipper to trim primer and adaptor regions from both ends
147 (https://github.com/agordon/fastx_toolkit). Sequences lacking a primer region on both ends of
148 the merged reads were discarded. Sequences were quality trimmed to have a maximum expected
149 number of errors per read of less than 0.1 and only sequences with more than 3 identical
150 replicates were included in downstream analyses. BLASTN 2.2.30+ was run locally, with a
151 representative sequence for each OTU as the query and the current NCBI nt nucleotide and
152 taxonomy database as the reference. The tabular BLAST hit tables for each OTU representative
153 were then parsed so only hits with > 97% query coverage and identity were kept.

Sequences were clustered into OTUs at the ≥ 97% sequence similarity level and sequence abundance counts for each OTU were determined using the usearch7 approach. The National Center for Biotechnology Information (NCBI) genus names associated with each hit were used to populate the OTU taxonomy assignment lists. Sequences that did not match over 90% of the query length and did not have at least 85% identity were considered unclassified. Otherwise the top BLASTn hit was used.

Statistical analyses

To quantify the accumulation of 23S OTUs with increasing numbers of samples, we used the *specaccum* function of the *vegan* package with the Lomolino function to describe the curves (Oksanen et al. 2017).

Hierarchical clustering of 23S was based on Ward's minimum variance method. A heat map was generated with the *heatmap.2* function of the *gplots* package (Warnes et al. 2016) using distance matrices created from the relative abundance of the top 50 23S OTUs. To identify taxa disproportionately associated with the 8 major clusters, indicator values were calculated for each of the top 50 OTUs based on abundance of occurrence (Dufrêne and Legendre 1997).

To assess the relationships between nutrient concentrations and 23S OTU abundance, forward stepwise regression was performed for [NH_4^+], [NO_3^-], and [PO_4^{3-}] with the top 50 23S OTUs ($P < 0.01$ for entry). To assess the relationships between phytoplankton richness and predictors, all singletons were removed from the abundance of reads, phytoplankton richness was first rarefied to the minimum number of reads for the sample set (4,212) and then regressed in a backwards elimination stepwise regression with nutrient concentration data, distance along the river, and secchi disk depth.

To compare 23S and 16S patterns, we restricted 16S data to the top 100 OTUs, representing 76% of the total reads. Previously, the 16S region was sequenced for the two particle size fractions independently (Henson et al. in review). Here, bacterial OTU abundance was averaged for the two fractions for a given sample. Mantel tests (*mantel* function of the *vegan* package) assessed Pearson correlations among assemblage similarity matrices, which were based on Euclidean distances. A cophenetic correlation was assessed for the 23S and 16S distance matrices using the *cor_cophenetic* function of *dendextend* package (Galili 2015). To visualize similarity in clustering between 23S and 16S OTU abundances, a tanglegram was generated using the *tanglegram* function of *dendextend* package based on the 23S hierarchical clustering and a new hierarchical clustering of 16S data also based on Ward's minimum variance method. The same stepwise regression technique on nutrient concentrations was used for the top 50 16S OTUs as was done for the 23S OTUs.

All statistical analyses were conducted in R 3.2.5 using Rstudio v. 1.0.136 except the stepwise regressions, which were computed in JMP v. 13.0.0 (SAS Institute, Cary NC, USA).

Results

Across all samples, the most abundant phytoplankton OTU was for taxa similar to *Thalassiosira rotula*, which represented 37.6% of all reads. The next most abundant OTU was for taxa similar to the diatom *Melosira tropica*, which represented 15.8% of all reads. In general, the top 10 OTUs represented 80.9% of all reads and the top 50 OTUs represented 96.4% of all reads. Among the top 50 OTUs, 72.6% of the reads were from Bacillariophyta, 8.7% were from Cryptophyta, and 7.0% were from Cyanobacteria. Chlorophyta and Eustigmatophyceae comprised 3.5% and 3.2% of the reads, respectively. Examining the pattern of OTU

accumulation, OTU abundance is predicted to asymptote at 447 OTUs with half of this occurring in 11.8 samples (Figure 1). Mean richness after rarefication was 55.3 ± 15.9 (s.d.) OTUs per sample. Mean richness increased at a rate of 7.1 ± 1.7 species per 1000 km ($r^2 = 0.23$, $P < 0.001$).

Phytoplankton had different patterns of distribution along the length of the river (Figure 2). Among the four most abundant OTUs, *Melosira tropica* OTU abundance peaked at approximately 60% of all reads 2750 km upstream from the Gulf of Mexico, while *Thalassiosira rotula* OTU abundance peaked at approximately 90% of all reads approximately 2250 km from the Gulf. In contrast, *Cyclotella* sp. WC03 (OTU 48) did not peak until ~1300 km from the Gulf (17% of all reads) and the *Skeletonema marinoi* OTU continued to increase below the confluence with the Missouri River, until it reached approximately 30% of the reads at the Gulf of Mexico. There were no general trends in the abundance of phytoplankton groups with respect to distance along the river when read abundance for the top 50 OTUs was aggregated by phylum (Figure 3).

Clustering of sites

The phytoplankton of Lake Itasca was the most unique set of OTUs and did not cluster with any other samples (Figure 4). The Lake Itasca assemblage was characterized by the abundance of chrysophyte species similar to *Ochromonas danica*, dinoflagellates similar to *Dinophysis fortii* and species similar to the yellow-green alga *Trachydiscus minutus* (Table 1). Beyond Lake Itasca, four other main clusters of sites were identified, which encompassed 57 of the remaining 61 samples. The first cluster contained 17 of the 27 samples taken upstream of the confluence with the Missouri River (Figure 4). These samples were indicated by their abundances of taxa similar in sequence to *Thalassiosira rotula* ($P = 0.003$; Table 1). The second cluster consisted of 8

221 samples in the Upper Mississippi that ranged along 300 km from Lake Pepin in Minnesota to
222 Dubuque, Iowa. These sites were indicated by their abundances of species similar in sequence to
223 the dinoflagellate *Gymnodinium eucyaneum*, the diatom *Tenuicylindrus* sp., the cryptomonad
224 *Cryptochloris*, and the diatom *Melosira tropica* (Table 1). The third main cluster denoted 20 of
225 the samples below the confluence with the Missouri River, primarily by their abundance of taxa
226 similar in sequence to *Skeletonema marinoi*. The fourth main cluster contained 12 samples below
227 the Missouri River confluence from above Vicksburg, MS to just below Three Rivers Wildlife
228 Management Area. These samples were indicated by their abundances of species similar in
229 sequence to the cryptomonads *Teleaulax acuta*, *Cryptomonas* sp., and *Plagioselmis*
230 *nannoplanctica* as well as two diatom OTUs for species similar in sequence to *Cyclotella* sp.
231 (Table 1).
232

233 **Relationships with nutrient data**

234 The best predictor of NH_4^+ concentrations at a given location was the abundance of diatoms
235 similar in sequence to *Sellaphora pupula*, which explained 38% of variation in [NH_4^+] (Table 2),
236 but mostly as a result of two sites having high [NH_4^+] (> 25 µg L^{-1}) and abundances of
237 *Sellaphora pupula*. After this OTU, the abundances of no other phytoplankton OTU predicted
238 NH_4^+ concentrations ($P > 0.01$ for all OTUs). [NO_3^-] was best predicted by 4 diatom OTUs,
239 which explained 61% of the variation in [NO_3^-]. Nitrate concentrations decreased with increasing
240 abundances of species similar in sequence to *Melosira tropica* and increased with increasing
241 abundances of species similar in sequence to *Cyclotella* sp. WC03, *Navicula salinicola*, and
242 *Dinophysis fortii* (Table 2). 80% of the variation in [PO_4^{3-}] was explained by the abundances of
243 six diatom OTUs (Table 2). [PO_4^{3-}] increased with increasing abundances of species similar in

244 sequence to *Skeletonema marinoi*, *Cyanobium* sp. *Navicula salinicola*, *Cyclotella* sp. WC03,
245 *Cryptomonas ovata*, and *Dinophysis fortii.* Phytoplankton OTU richness increased downstream
246 ($P < 0.01$), but in the backwards elimination stepwise regression, phytoplankton OTU richness
247 (intercept = 16.42 ± 6.83; $P = 0.02$) increased with increasing secchi disk depth (0.295 ± 0.071
248 OTUs cm^{-1}; $P < 0.001$) and with increasing $[PO_4^{3-}]$ (0.328 ± 0.048 OTUs $(\mu g\ L^{-1})^{-1}$; $P < 0.001$)
249 (Figure 5).

250 **Comparing phytoplankton and bacteria**

251 Comparing distance matrices with a Mantel test, 23S and 16S assemblages were correlated ($r =$
252 0.44, $P < 0.001$). Similarly, the hierarchical clustering of sites based on 23S and 16S
253 assemblages were correlated (cophenetic correlation, $r = 0.43$), revealing structural similarity in
254 the two assemblages. For example, comparing the dendrograms, paired samples often clustered
255 together for both the 23S and 16S assemblages, such as the D samples and Aa samples. Also,
256 sites U and W were more similar to one another than other sites for both 23S and 16S (Figure 6).
257 Stepping back to the broader patterns, the major clusters of sites in the 23S data were also largely
258 present for the 16S data, though the relative positions within this cluster were mixed. Some
259 differences in the clustering between the two sets of samples were likely due to stochasticity or
260 contamination in individual samples for one primer pair. For example, with the 23S data, site P
261 clustered with the Al sites. Yet, in the 16S data it clustered more closely with the adjacent O sites.

262 Compared to phytoplankton, using the same forward stepwise regression technique—top 50
263 OTUs, $P < 0.01$ for entry—bacterial OTUs typically explained a greater proportion of nutrient
264 concentrations in the water. For $[NH_4^+]$, five bacterial OTUs predicted 69% of the variation in
265 $[NH_4^+]$ compared to 38% of the variation with phytoplankton (Table 3). Sites with greater
266 abundances of three OTUs (a Firmicutes, a Bacteroidetes, and a Proteobacteria) had higher

267 [NH_4^+] while sites with greater abundances of two OTUs (an Actinobacteria and a Bacteriodetes)
268 had lower [NH_4^+] (Table 3). For [NO_3^-], phytoplankton had predicted 61% of the variation, but
269 the abundances of six bacterial OTUs explained 80%. [NO_3^-] concentrations increased with
270 increasing abundances of two bacterial OTUs (an Actinobacteria and a Bacteriodetes) and
271 decreased with increasing abundances of four bacterial OTUs (an Actinobacteria, a Bacteriodetes,
272 and two Proteobacteria) (Table 3). For [PO_4^{3-}], bacteria predicted 81% of the variation in
273 concentrations, compared to 80% for phytoplankton. [PO_4^{3-}] were lower with increasing
274 abundances of three bacterial OTUs (an Actinobacteria, a Bacteriodetes, and a Proteobacteria)
275 and increased with increasing abundances of a Planctomycetes OTU (Table 3).

276 ## Discussion

277 Overall, this research demonstrates the potential of sequencing the 23S region in water samples
278 to reconstruct a broad diversity of the phytoplankton assemblage and provide information on
279 underlying environmental conditions. Here, we saw that 23S-derived phytoplankton assemblages
280 shifted along the length of the river, paralleled shifts in bacterial assemblages, and could predict
281 abiotic conditions such as aquatic inorganic nutrient concentrations. These results support the
282 further development of 23S sequencing of aquatic eDNA to reconstruct phytoplankton
283 assemblages in order to infer environmental conditions.

284 The patterns in phytoplankton eDNA abundance observed for the Mississippi River were similar
285 to those in other rivers. For example, Cannon et al. sequenced both 16S and 23S along the length
286 of the Cuyahoga River in northern Ohio, USA (Cannon et al. 2017). As with the Mississippi, in
287 the Cuyahoga, phytoplankton 23S OTUs were spatially patterned and many phytoplankton and
288 bacteria were correlated along the length of the river, potentially reflecting underlying

289 environmental conditions. In another study, Craine et al. sequenced 4 primer pairs including 16S
290 and 23S to reconstruct biotic assemblages along 475 km of the Potomac River in Maryland, USA
291 (Craine et al. in review). As with the Mississippi River, phytoplankton assemblages were
292 distinctly patterned along the river and were strongly associated with river size and aquatic
293 phosphorus concentrations. For the Potomac, phytoplankton richness increased downstream, just
294 as with the Mississippi. Among these three studies, there were strong differences in
295 phytoplankton assemblages. For example, compared to the Mississippi River, the Cuyahoga
296 River had a greater dominance of Cryptophytes. Although both the lower Potomac and
297 Mississippi were dominated by diatoms, different diatoms dominated the two rivers.

298 Traditional sampling of large rivers with visual quantification of phytoplankton also showed
299 many similar patterns as we observed here. For example, in the River Loire, large portions of the
300 river were dominated by diatoms and many taxa were associated with eutrophic conditions
301 (Abonyi et al. 2012). In the Upper Missouri/Mississippi/Ohio River basin in 2004/5,
302 phytoplankton diatom assemblages responded to agricultural disturbance, urbanization, and
303 eutrophication (Kireta et al. 2012b). Compared to the other two rivers, the upper Mississippi
304 River was distinguished by its high levels of eutrophication, with many of the taxa observed here
305 in high abundance (or congeners) being indicative of eutrophic and/or high agricultural or urban
306 disturbance (Kireta et al. 2012b). The lower Mississippi River is also considered generally
307 eutrophic and many of the taxa that indicated eutrophic or saline conditions were similar to those
308 that dominated assemblages here (Bellinger et al. 2013).

309 Empirically, given the greater abundance of phytoplankton in waters than, for example, insects
310 or fish, there has been greater success sequencing the eDNA of phytoplankton than larger
311 organisms. This further favors developing the use of phytoplankton over other taxa. The ability

in this study of the relative abundance of phytoplankton to predict aquatic nutrient concentrations should encourage future research to develop this technique for bioassessment. For example, sections of the Mississippi River with high abundances of *Melosira tropica* and *Navicula salinicola* or low abundances of a *Cyclotella* OTU had high [NO_3^-]. If these relationships were to hold up across different river systems and seasons, then the abundances of these species as determined by sequencing eDNA could broadly serve as an indicator of [NO_3^-] without having to measure it directly. Given that using a similar technique, strong relationships between aquatic nutrient concentrations and phytoplankton abundances were seen in the Potomac River, too (Craine et al. in review), this method continues to show promise as a bioassessment tool. To our knowledge, there is no theory to explain why phytoplankton diversity increases with distance downstream and/or with increased [PO_4^{3-}], which was also observed in the Potomac. In fact, such observations actually run counter to the popular River Continuum Concept, which postulated that after an initial increase in headwaters, diversity should decrease with increasing river size (Vannote et al. 1980). However our observed trend of increasing diversity is consistent with other measurements of Mississippi River microbial assemblages (Payne et al. 2017)(Henson et al. in review). Theory aside, it will take much larger datasets to assess whether phytoplankton diversity, in and of itself, is diagnostic for any environmental conditions.

Greater taxonomic resolution is likely possible with other primer pairs in conjunction with 23S, but there is no evidence yet that this is necessary. That said, there are still areas where more research is required before metabarcoding with 23S for phytoplankton assemblages can be operationalized. For example, the number of sequences known from diatom taxa is small fraction of the several thousand species described from North America (Kociolek 2006). Of the nearly one thousand taxa listed in the Diatoms of the US web flora (Spaulding et al. 2010),

335 approximately one hundred have associated sequences that are currently available in GenBank.
336 Other studies (Visco et al. 2015) report that only 28% of taxa identified by microscopy had
337 corresponding reads in sequence data. Consequently, for diatoms, the OTUs were mapped to taxa
338 that were most similar, with the outcome that species that are well-characterized in gene
339 sequences (i.e. *Melosira tropica*) that have not reported from inland waters in river surveys (U.S.
340 Geological Survey BioData).

341 It is possible that the OTU matches to *Melosira tropica* could reflect the presence of the very
342 common *M. varians. Melosira tropica* has not been reported from inland waters, but *M. varians*
343 is one of the very common river species (Potapova and Charles 2007). Although it is not
344 expected to find *Thalassiosira rotula* in the more northern reaches of the Mississippi River,
345 others (Visco et al. 2015) report that the common *Stephanodiscus minutulus* was included in a
346 well-supported clade with a number of *Thalassiosira* species, at least based on the particular
347 region examined. *Skeletonema potamos* and *S. costatum* are commonly reported from rivers with
348 high conductivity, resulting from agricultural input (Potapova and Charles 2007). For example,
349 both of these species have been found in national surveys in rivers including the Milwaukee
350 River at Milwaukee WI and the Maumee River at Waterville OH.

351 Beyond improving reference databases, autecological information for many phytoplankton taxa
352 exist (Reynolds et al. 2002, Padisák et al. 2009), but indices generated with 23S will need to
353 continue to be calibrated against environmental conditions with multiple reference sites to ensure
354 that there are not covariates driving the relationship. For example, nutrient concentrations,
355 distance downstream, and time of sampling were all associated in this study and these other
356 factors could be influencing the relationships we observed between phytoplankton assemblages
357 and nutrient availability. Multiple large rivers of different nutrient status will need to be included

to partition out the direct effects of nutrient concentrations from other covariates driving assemblage composition. Reference databases will also need to be expanded by sequencing a larger diversity of phytoplankton organisms and identifying taxa associated with sequences. Although eDNA-based bioassessment can occur independent of taxonomy (Apotheloz-Perret-Gentil et al. 2017), more robust, stable indices will likely require ecological information about individual taxa, too. Given the breadth of taxa sequenced with 23S, this means broad biodiversity surveys are required for all phytoplankton taxa rather than a single taxonomic group, such as cyanobacteria. Although this technique should work with periphyton also, future work should continue to test whether phytoplankton or periphyton are best for bioassessment of given environmental conditions (Kireta et al. 2012a), although previous work with traditional techniques appears to favor the utility of phytoplankton over periphyton for assessing environmental conditions in some large rivers (Reavie et al. 2010), though not others (Bellinger et al. 2013).

372 ## Acknowledgments

373 This work was supported in part by the Department of Biological Sciences, College of Science,
374 and the Office of Research and Economic Development at Louisiana State University; and the
375 College of the Environment at the University of Washington. The authors appreciate comments
376 from Kristy Deiner and Ian Bishop, Meredith Tyree and Nicholas Schulte for discussions on data
377 interpretation. Sequence data are available at
378 www.ncbi.nlm.nih.gov/Traces/study/?acc=SRP132323.

379

380

References

Abonyi, A., M. Leitão, A. M. Lançon, and J. Padisák. 2012. Phytoplankton functional groups as indicators of human impacts along the River Loire (France). Hydrobiologia **698**:233-249.

American Sportfishing Association. 2015. Economic Contributions of Recreational Fishing: U.S. Congressional Districts.

Angradi, T. R., D. W. Bolgrien, T. M. Jicha, M. S. Pearson, B. H. Hill, D. L. Taylor, E. W. Schweiger, L. Shepard, A. R. Batterman, M. F. Moffett, C. M. Elonen, and L. E. Anderson. 2009. A bioassessment approach for mid-continent great rivers: the Upper Mississippi, Missouri, and Ohio (USA). Environmental Monitoring and Assessment **152**:425-442.

Apotheloz-Perret-Gentil, L., A. Cordonier, F. Straub, J. Iseli, P. Esling, and J. Pawlowski. 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. Mol Ecol Resour.

Barbour, M. T., J. Gerritsen, B. D. Snyder, and J. B. Stribling. 1999. Rapid bioassessment protocols for use in streams and wadeable rivers: periphyton, benthic macroinvertebrates and fish. US Environmental Protection Agency, Office of Water Washington, DC.

Barnes, M. A., and C. R. Turner. 2015. The ecology of environmental DNA and implications for conservation genetics. Conservation Genetics **17**:1-17.

Bellinger, B. J., T. R. Angradi, D. W. Bolgrien, T. M. Jicha, B. H. Hill, and E. D. Reavie. 2013. Longitudinal variation and response to anthropogenic stress in diatom assemblages of the Lower Mississippi River, USA. River Systems **21**:29-54.

Buss, D. F., D. M. Carlisle, T.-S. Chon, J. Culp, J. S. Harding, H. E. Keizer-Vlek, W. A. Robinson, S. Strachan, C. Thirion, and R. M. Hughes. 2015. Stream biomonitoring using

macroinvertebrates around the globe: a comparison of large-scale programs. Environmental Monitoring and Assessment **187**:1.

Cannon, M. V., J. Craine, J. Hester, A. Shalkhauser, E. R. Chan, K. Logue, S. Small, and D. Serre. 2017. Dynamic microbial populations along the Cuyahoga River. PLOS ONE **12**:e0186290.

Cannon, M. V., J. Hester, A. Shalkhauser, E. R. Chan, K. Logue, S. T. Small, and D. Serre. 2016. In silico assessment of primers for eDNA studies using PrimerTree and application to characterize the biodiversity surrounding the Cuyahoga River. Sci Rep **6**:22908.

Craine, J. M., M. V. Cannon, A. J. Elmore, S. M. Guinn, and N. Fierer. in review. DNA metabarcoding potentially reveals multi-assemblage eutrophication responses in an eastern North American river. PLOS ONE.

Davies, S. P., and S. K. Jackson. 2006. The biological condition gradient: a descriptive model for interpreting change in aquatic ecosystems. Ecological Applications **16**:1251-1266.

Deiner, K., E. A. Fronhofer, E. Machler, J. C. Walser, and F. Altermatt. 2016. Environmental DNA reveals that rivers are conveyer belts of biodiversity information. Nat Commun **7**:12544.

Dufrêne, M., and P. Legendre. 1997. Species assemblages and indicator species: the need for a flexible asymmetrical approach. Ecological Monographs **67**:345-366.

Galili, T. 2015. dendextend: an R package for visualizing, adjusting, and comparing trees of hierarchical clustering. Bioinformatics.

Goffaux, D., G. Grenouillet, and P. Kestemont. 2005. Electrofishing versus gillnet sampling for the assessment of fish assemblages in large rivers. Archiv für Hydrobiologie **162**:73-90.

426 Hamsher, S. E., M. M. LeGresley, J. L. Martin, and G. W. Saunders. 2013. A comparison of
427 morphological and molecular-based surveys to estimate the species richness of
428 Chaetoceros and Thalassiosira (bacillariophyta), in the Bay of Fundy. PLoS One
429 **8**:e73521.
430 Jackson, C. R., J. J. Millar, J. T. Payne, and C. A. Ochs. 2014. Free-Living and Particle-
431 Associated Bacterioplankton in Large Rivers of the Mississippi River Basin Demonstrate
432 Biogeographic Patterns. Appl Environ Microbiol **80**:7186-7195.
433 Karr, J. R. 1999. Defining and measuring river health. Freshwater Biology **41.2**:221-234.
434 Kelly, M., and B. Whitton. 1995. The trophic diatom index: a new index for monitoring
435 eutrophication in rivers. Journal of Applied Phycology **7**:433-444.
436 Kireta, A. R., E. D. Reavie, G. V. Sgro, T. R. Angradi, D. W. Bolgrien, B. H. Hill, and T. M.
437 Jicha. 2012a. Planktonic and periphytic diatoms as indicators of stress on great rivers of
438 the United States: Testing water quality and disturbance models. Ecological Indicators
439 **13**:222-231.
440 Kireta, A. R., E. D. Reavie, G. V. Sgro, T. R. Angradi, D. W. Bolgrien, T. M. Jicha, and B. H.
441 Hill. 2012b. Assessing the condition of the Missouri, Ohio, and Upper Mississippi rivers
442 (USA) using diatom-based indicators. Hydrobiologia **691**:171-188.
443 Lyon, J. P., T. Bird, S. Nicol, J. Kearns, J. O'Mahony, C. R. Todd, I. G. Cowx, C. J. A.
444 Bradshaw, and J. M. Jech. 2014. Efficiency of electrofishing in turbid lowland rivers:
445 implications for measuring temporal change in fish populations. Canadian Journal of
446 Fisheries and Aquatic Sciences **71**:878-886.

447 Mächler, E., K. Deiner, P. Steinmann, and F. Altermatt. 2014. Utility of environmental DNA for
448 monitoring rare and indicator macroinvertebrate species. Freshwater Science **33**:1174-
449 1183.
450 Malmqvist, B., and S. Rundle. 2002. Threats to the running water ecosystems of the world.
451 Environmental Conservation **29**:134-153.
452 Oksanen, J., F. G. Blanchet, M. Friendly, R. Kindt, P. Legendre, D. McGlinn, P. R. Minchin, R.
453 B. O'Hara, G. L. Simpson, P. Solymos, M. H. H. Stevens, E. Szoecs, and H. Wagner.
454 2017. vegan: Community Ecology Package. R package version 2.4-3.
455 Olds, B. P., C. L. Jerde, M. A. Renshaw, Y. Li, N. T. Evans, C. R. Turner, K. Deiner, A. R.
456 Mahon, M. A. Brueseke, P. D. Shirey, M. E. Pfrender, D. M. Lodge, and G. A. Lamberti.
457 2016. Estimating species richness using environmental DNA. Ecol Evol **6**:4214-4226.
458 Padisák, J., L. O. Crossetti, and L. Naselli-Flores. 2009. Use and misuse in the application of the
459 phytoplankton functional classification: a critical review with updates. Hydrobiologia
460 **621**:1-19.
461 Palaniappan, M., P. Gleick, L. Allen, M. Cohen, J. Christian-Smith, and C. Smith. 2010.
462 Clearing the waters: A focus on water quality solutions. Nairobi, Kenya: UNEP/Pacific
463 Institute.
464 Payne, J. T., J. J. Millar, C. R. Jackson, and C. A. Ochs. 2017. Patterns of variation in diversity
465 of the Mississippi river microbiome over 1,300 kilometers. PLoS One **12**:e0174890.
466 Potapova, M., and D. F. Charles. 2007. Diatom metrics for monitoring eutrophication in rivers of
467 the United States. Ecological Indicators **7**:48-70.

468 Reavie, E. D., T. M. Jicha, T. R. Angradi, D. W. Bolgrien, and B. H. Hill. 2010. Algal
469 assemblages for large river monitoring: Comparison among biovolume, absolute and
470 relative abundance metrics. Ecological Indicators **10**:167-177.
471 Reyjol, Y., G. Loot, and S. Lek. 2005. Estimating sampling bias when using electrofishing to
472 catch stone loach. Journal of Fish Biology **66**:589-591.
473 Reynolds, C. S., V. Huszar, C. Kruk, L. Naselli-Flores, and S. Melo. 2002. Towards a functional
474 classification of the freshwater phytoplankton. Journal of plankton research **24**:417-428.
475 Schoolmaster, D. R., J. B. Grace, and E. William Schweiger. 2012. A general theory of
476 multimetric indices and their properties. Methods in Ecology and Evolution **3**:773-781.
477 Stein, E. D., M. C. Martinez, S. Stiles, P. E. Miller, and E. V. Zakharov. 2014a. Is DNA
478 barcoding actually cheaper and faster than traditional morphological methods: results
479 from a survey of freshwater bioassessment efforts in the United States? PLoS One
480 **9**:e95525.
481 Stein, E. D., B. P. White, R. D. Mazor, J. K. Jackson, J. M. Battle, P. E. Miller, E. M. Pilgrim,
482 and B. W. Sweeney. 2014b. Does DNA barcoding improve performance of traditional
483 stream bioassessment metrics? Freshwater Science **33**:302-311.
484 Thomsen, P. F., and E. Willerslev. 2015. Environmental DNA – An emerging tool in
485 conservation for monitoring past and present biodiversity. Biological Conservation
486 **183**:4-18.
487 US Environmental Protection Agency. 2015. A compilation of cost data associated with the
488 impacts and control of nutrient pollution. US EPA Office of Water.
489 Vannote, R. L., G. W. Minshall, K. W. Cummins, J. R. Sedell, and C. E. Cushing. 1980. The
490 river continuum concept. Canadian Journal of Fisheries and Aquatic Sciences **37**:130-137.

Warnes, G. R., B. Bolker, L. Bonebakker, R. Gentleman, W. Huber, A. Liaw, T. Lumley, M. Maechler, A. Magnusson, S. Moeller, M. Schwartz, and B. Venables. 2016. gplots: Various R Programming Tools for Plotting Data. R package version 3.0.1.

Young, R. A., and J. B. Loomis. 2014. Determining the economic value of water: concepts and methods. Routledge.

Zimmermann, J., G. Glockner, R. Jahn, N. Enke, and B. Gemeinholzer. 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Mol Ecol Resour **15**:526-542.

503 Table 1. Table of indicator values for different clusters. See Figure 4 for which sites belong to
504 each cluster.

OTU	Taxon	Cluster	Indicator value	*P* value
OTU.12	*Planktothrix agardhii*	Site A	0.85	<0.001
OTU.29	*Cryptomonas obovoidea*	Site A	0.83	0.03
OTU.79	*Skeletonema marinoi*	2	0.74	<0.001
OTU.10	*Teleaulax acuta*	3	0.66	<0.001
OTU.37	*Cryptomonas* sp. Sinjeong 080610A	3	0.61	0.02
OTU.338	*Plagioselmis nannoplanctica*	3	0.41	0.03
OTU.192	*Cyclotella sp.* WC03_2	3	0.36	<0.001
OTU.48	*Cyclotella sp*. WC03_2	3	0.34	0.002
OTU.49	*Choricystis parasitica*	Sample Ai1	0.99	0.003
OTU.418	*Nannochloropsis salina*	Sample Ai1	0.92	0.01
OTU.77	*Navicula salinicola*	Sample Ai1	0.58	0.006
OTU.96	*Gymnodinium eucyaneum*	Sample Ai1	0.55	0.04
OTU.1	*Thalassiosira rotula*	5	0.30	0.003
OTU.145	*Gymnodinium eucyaneum*	6	0.67	0.03
OTU.189	*Tenuicylindrus* sp. LG-2015	6	0.65	<0.001
OTU.100	*Cryptochloris* sp. PR-2015	6	0.62	0.04
OTU.3	*Melosira tropica*	6	0.43	0.002
OTU.43	*Ochromonas danica*	Itasca	1.00	0.05
OTU.28	*Dinophysis fortii*	Itasca	0.98	0.01
OTU.14	*Trachydiscus minutus*	Itasca	0.87	<0.001

OTU.50	*Synechococcus* sp. RCC307	Samples U1,W1	0.50	0.04

505

506

507 Table 2. 23S OTU predictors of nutrient concentrations including sums of squares, estimates (µg
508 L^{-1}), and *P* values. Coefficients of variation for [NH_4^+], [NO_3^-], and [PO_4^{3-}] were 0.38, 0.61, 0.80,
509 respectively.

510

Nutrient	Variable	%SS	Estimate	*P* value
[NH_4^+]	Intercept		3.2 ± 0.8	<0.001
	Sellaphora pupula	100%	492.4 ± 75.7	<0.001
[NO_3^-]	Intercept		1046 ± 148.4	<0.001
	Cyclotella sp. WC03_2	34.1%	-1947.5 ± 453.1	<0.001
	Melosira tropica	23.0%	4152.7 ± 1177.9	<0.001
	Navicula salinicola	22.6%	67288.8 ± 19245.2	<0.001
	Dinophysis fortii	20.3%	5803301 ± 1749619.8	0.002
[PO_4^{3-}]	Intercept		2.4 ± 6.4	0.7095
	Cyclotella sp. WC03_2	8.2%	248.3 ± 61.1	<0.001
	Skeletonema marinoi	38.2%	317.3 ± 36.2	<0.001
	Cyanobium sp. PCC 7009	27.6%	352.2 ± 47.2	<0.001
	Cryptomonas ovata	9.9%	1597.8 ± 357.6	<0.001
	Navicula salinicola	9.9%	3977.9 ± 891.1	<0.001
	Dinophysis fortii	6.2%	299742.4 ± 84789.6	<0.001

511

512

513

514

515 Table 3. 16S OTU predictors of nutrient concentrations including sums of squares, estimates (µg
516 L^{-1}), and *P* values. Coefficients of variation for [NH_4^+], [NO_3^-], and [PO_4^{3-}] were r^2 = 0.69, 0.78,
517 0.81, respectively.

Nutrient	Variable	%SS	Estimate	*P* value
[NH_4^+]	Intercept		0.2 ± 1.3	0.84
	OTU13	21.2%	167.8 ± 36.7	<.001
	OTU17	28.8%	-133.1 ± 25	<.001
	OTU15	9.2%	-52.5 ± 17.4	0.004
	OTU25	8.2%	138.0 ± 48.4	0.006
	OTU45	32.6%	899.3 ± 158.6	<.001
[NO_3^-]	Intercept		1875.2 ± 125.5	<.001
	OTU3	11.1%	6966.0 ± 1961.3	<0.001
	OTU7	11.7%	-14045.1 ± 3850.6	<0.001
	OTU19	9.6%	-49899.6 ± 15098.3	0.002
	OTU44	18.5%	-37316.1 ± 8125.9	<0.001
	OTU25	31.0%	32079.2 ± 5393.8	<0.001
	OTU35	18.2%	-20507.3 ± 4508.4	<0.001
[PO_4^{3-}]	Intercept		102.5 ± 8.7	<0.001
	OTU15	55.0%	-916.7 ± 105.6	<0.001
	OTU23	7.8%	-1054.3 ± 322.7	0.002
	OTU30	11.7%	3924.4 ± 981.9	<0.001
	OTU49	25.5%	-3925.7 ± 663.6	<0.001

518

519

520

521

522 Figure captions.

523 Figure 1. Accumulation of OTUs with additional samples. 50% of the accumulation of OTUs
524 occurs with 11.8 samples and richness is predicted to asymptote at 447 samples.

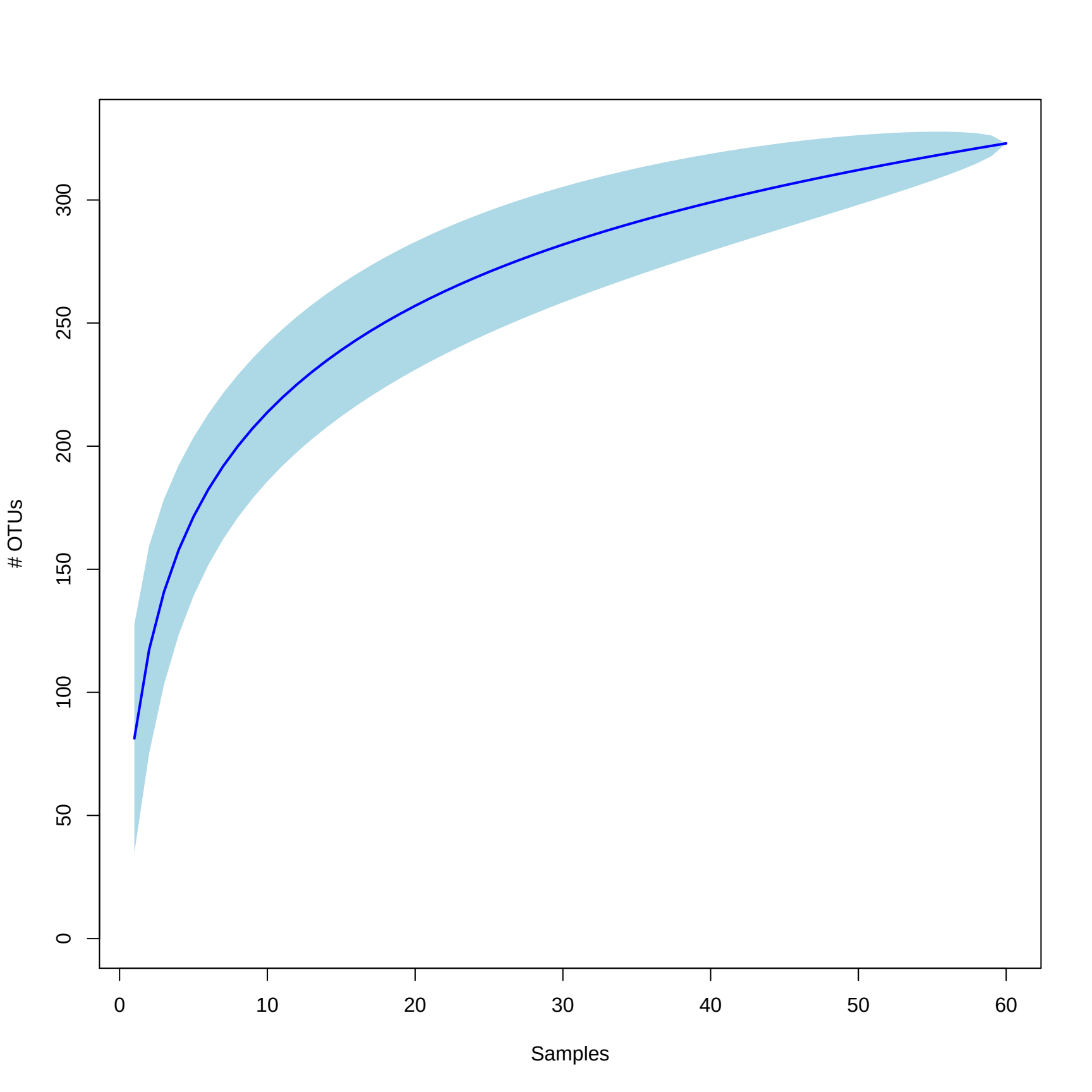
525 Figure 2. Relative read abundance of four most abundant OTUs as a function of distance from
526 the mouth of the Mississippi River.

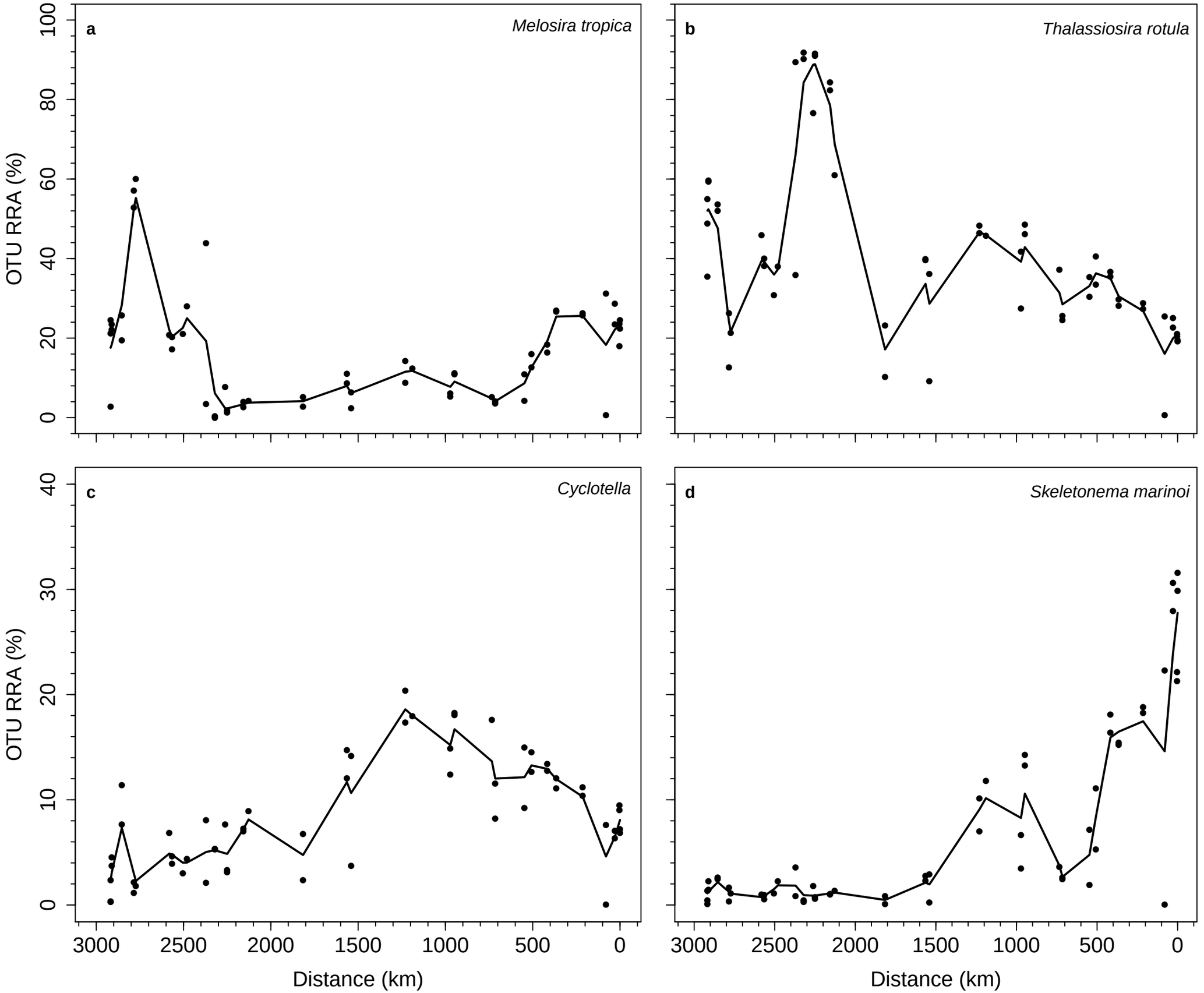
527 Figure 3. Relative read abundance of five main taxonomic groups as a function of distance from
528 the mouth of the Mississippi River.

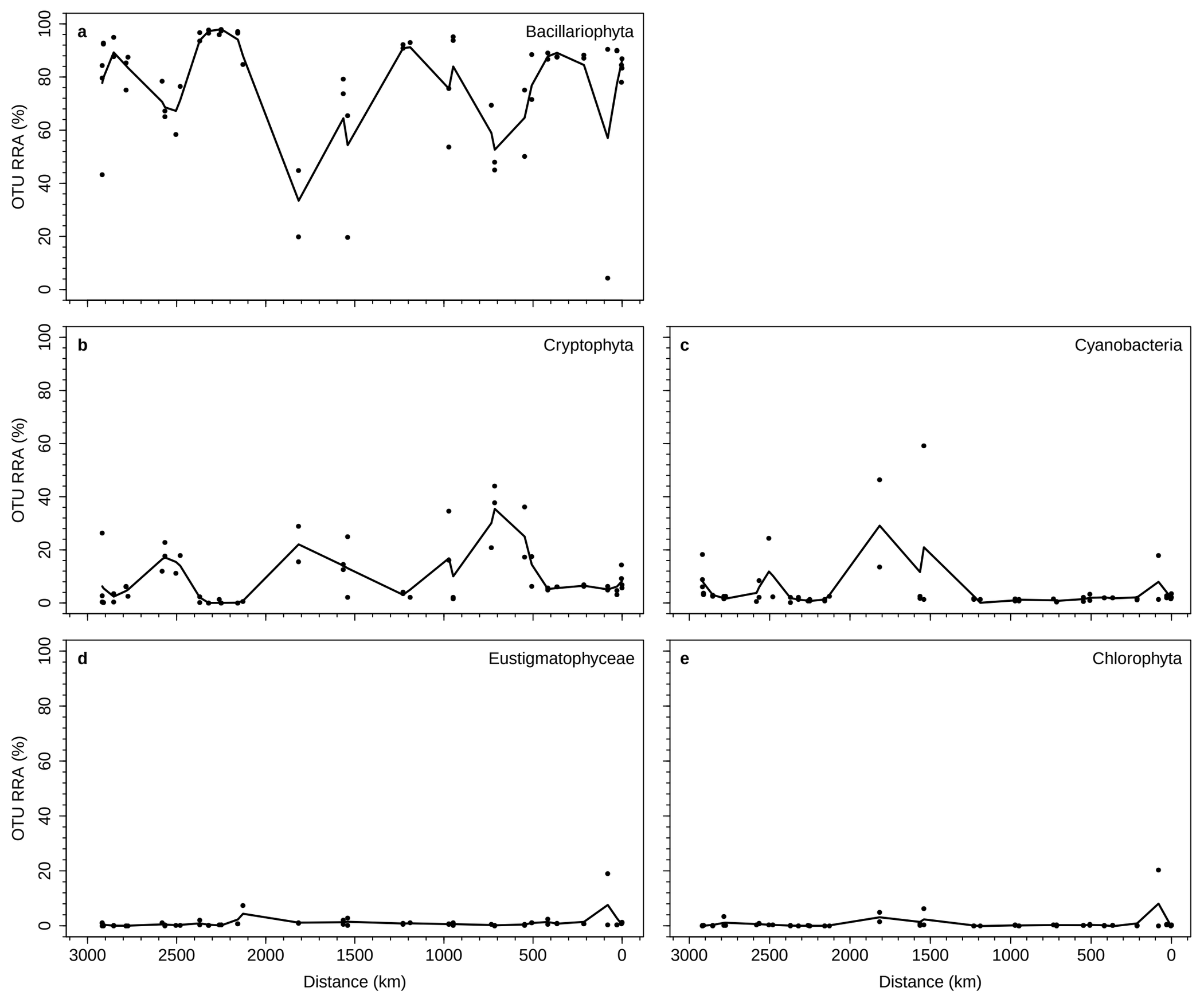
529 Figure 4. Heat map of abundances of OTUs at sites along the Mississippi River based on the
530 standardized relative abundance of the 50 most abundant 23S OTUs. Blue indicates a low
531 relative abundance and red high with gray intermediate. Sites and OTUs were clustered
532 hierarchically based on dissimilarity index of relative abundances. Four major site clusters
533 shown in color including Cluster 2 (purple), Cluster 5(green), Cluster 6 (red), and Cluster 3
534 (blue).

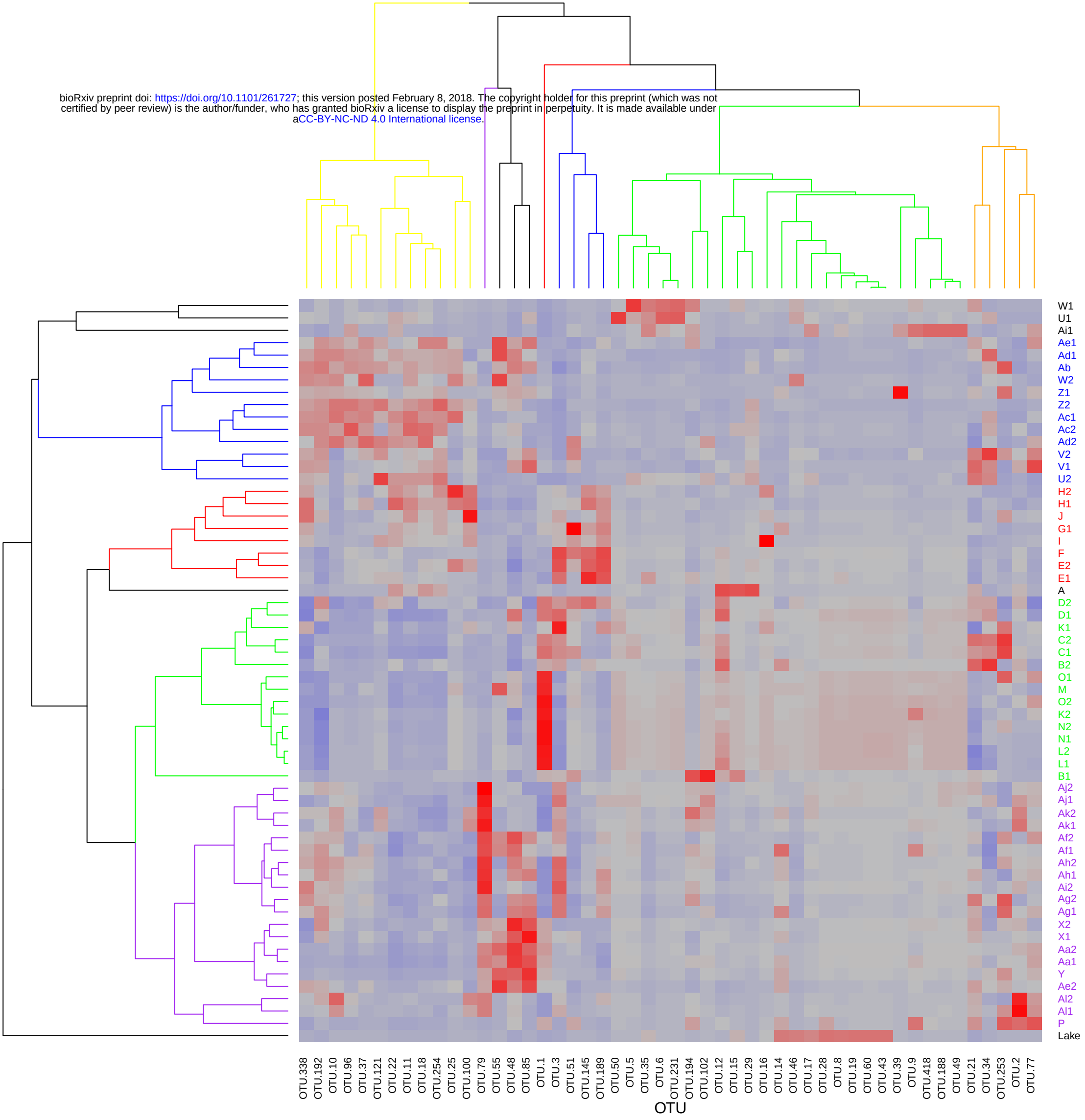
535 Figure 5. Partial residual plots of rarefied OTU richness as a function of (a) secchi disk depth and
536 (b) [PO_4^{3-}]. Non-significant variables include distance down the river, [NO_3^-], and [NH_4^+].

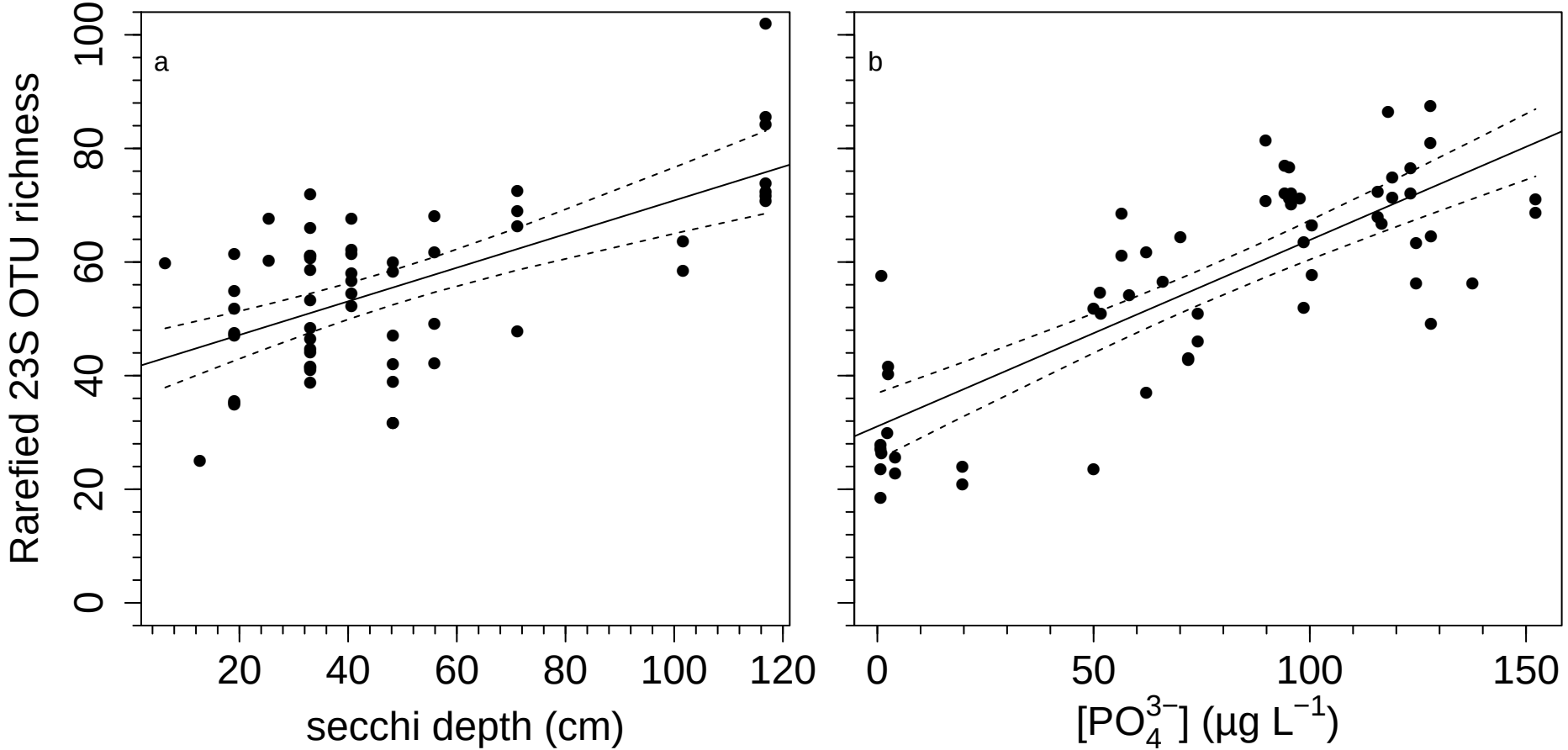
537 Figure 6. Tanglegram for the association between site hierarchical clusterings based on 23S and
538 16S OTU abundance. Colored lines between dendrogram tips represent similar relative
539 placement of sites within the clustering diagram.

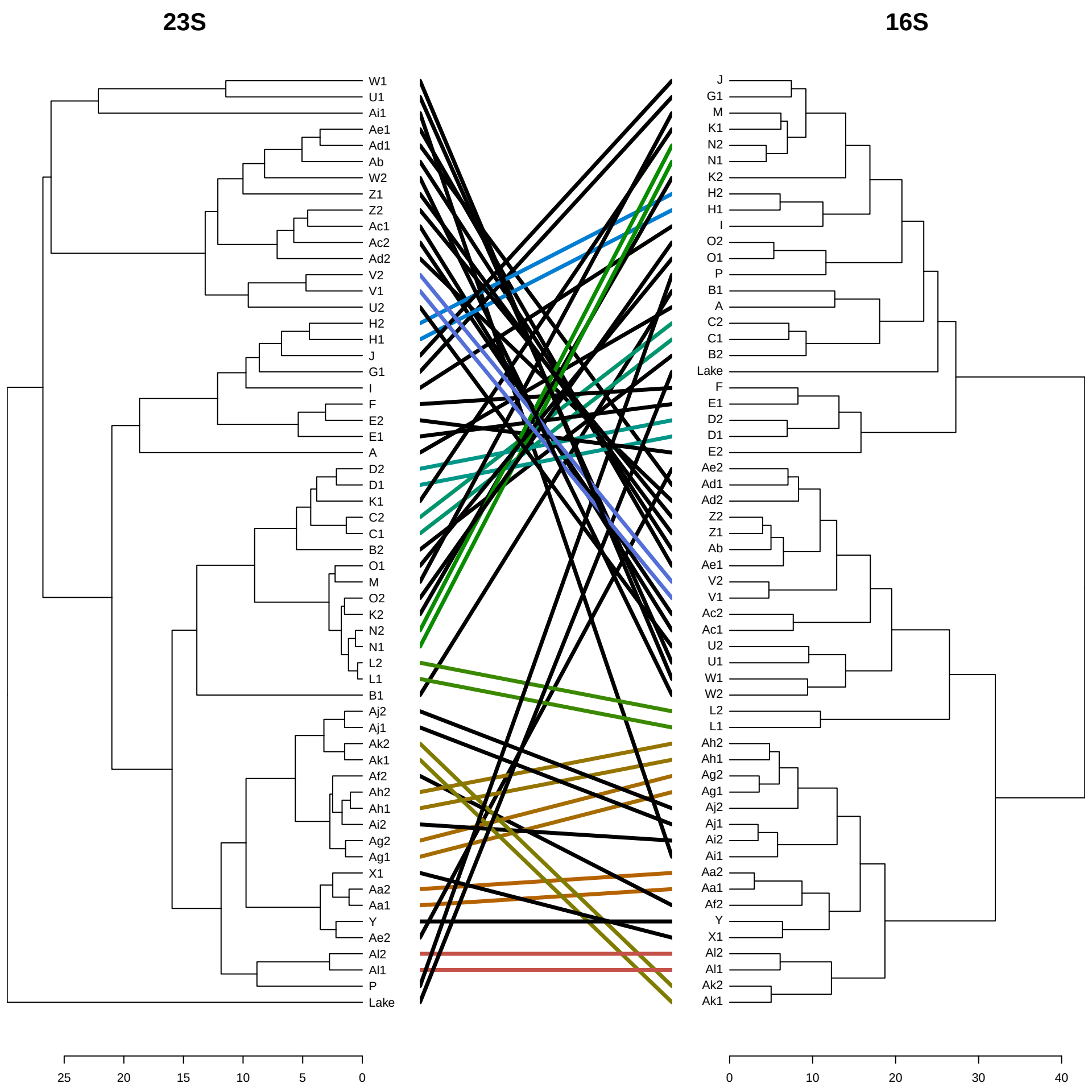

0
50
100
150
200
250
300
OTUs
0
10
20
30
40
50
60
Samples


a
Melosira tropica
b
Thalassiosira rotula
c
Cyclotella
d
Skeletonema marinoi
OTU RRA (%)
OTU RRA (%)
0
20
40
60
80
100
0
10
20
30
40
3000
2500
2000
1500
1000
500
0
Distance (km)
Distance (km)


a
Bacillariophyta
b
Cryptophyta
c
Cyanobacteria
d
Eustigmatophyceae
e
Chlorophyta
OTU RRA (%)
0
20
40
60
80
100
3000
2500
2000
1500
1000
500
0
Distance (km)




a
b
Rarefied 23S OTU richness
0
20
40
60
80
100
20
40
60
80
100
120
secchi depth (cm)
0
50
100
150
[PO43−] (µg L−1)


23S
16S
W1
U1
Ai1
Ae1
Ad1
Ab
W2
Z1
Z2
Ac1
Ac2
Ad2
V2
V1
U2
H2
H1
J
G1
I
F
E2
E1
A
D2
D1
K1
C2
C1
B2
O1
M
O2
K2
N2
N1
L2
L1
B1
Aj2
Aj1
Ak2
Ak1
Af2
Ah2
Ah1
Ai2
Ag2
Ag1
X1
Aa2
Aa1
Y
Ae2
Al2
Al1
P
Lake
J
G1
M
K1
N2
N1
K2
H2
H1
I
O2
O1
P
B1
A
C2
C1
B2
Lake
F
E1
D2
D1
E2
Ae2
Ad1
Ad2
Z2
Z1
Ab
Ae1
V2
V1
Ac2
Ac1
U2
U1
W1
W2
L2
L1
Ah2
Ah1
Ag2
Ag1
Aj2
Aj1
Ai2
Ai1
Aa2
Aa1
Af2
Y
X1
Al2
Al1
Ak2
Ak1
25
20
15
10
5
0
0
10
20
30
40