Pilot project: Evaluation of DNA-based metabarcoding of microeukaryotic assemblages for water-quality monitoring in New Jersey rivers and streams

FINAL REPORT

Prepared for:

New Jersey Department of Environmental Protection

Mihaela Enache

428 E. State Street

Trenton, NJ 08625

Prepared by:

Academy of Natural Sciences of Drexel University

Marina Potapova and Alison Minerovic

1900 Benjamin Franklin Parkway

Philadelphia, PA 19103

January 23, 2019

TABLE OF CONTENTS

Page

LIST OF TABLES

LIST OF FIGURES

Executive Summary

DNA metabarcoding is a rapidly expanding area of research and a novel approach to assessment of biodiversity, especially valuable for characterization of microbial communities. New studies investigating potential uses of metabarcoding for monitoring bacterial, algal, protozoan and invertebrate communities are being published at ever increasing rate. The benefits of metabarcoding include the ability to capture much greater biodiversity in comparison to visual/microscopic assessments, process a great number of samples simultaneously using the high-throughput next-generation sequencing technology, and to use genetic markers for identification purposes thus excluding the analyst bias. As the cost of sequencing continuously goes down and new approaches are being explored, it becomes more and more evident that metabarcoding will eventually be adopted as a tool for routine biomonitoring of aquatic systems. The current metabarcoding methods have not been developed yet to the point of producing highly accurate, reproducible and quantitatively unbiased assessments and the success rates have been variable among various studies. This why is it important to conduct studies aimed at developing appropriate methods for specific ecosystems and taxonomic groups of organisms.

The goal of this pilot project was to explore the DNA metabarcoding approach to biological assessment of New Jersey wadeable streams. We used methodology previously recommended to characterize the whole eukaryotic community by Illumina Mi-Seq sequencing of the V9 fragment of 18S rDNA gene following the Earth Microbiome Project protocol. In order to compare the outcome of molecular characterization of stream biofilms community with traditional microscopy-based assessment, we also enumerated diatoms in the same samples. To evaluate the response of biofilm communities to impairment we sampled 14 sites along a disturbance gradient. To estimate variability of the results obtained by both methods, we analyzed three field replicate samples per sampling site and two lab replicates per sample, which were two separate DNA extractions.

Metabarcoding revealed a vast variety of eukaryotic organisms inhabiting stream biofilms, including some sequences representing potentially pathogenic taxa. A total of 6,995,104 sequences that belonged to 5,869 unique Operational Taxonomic Units (OTUs) were recovered. Diatoms were the most abundant microeukaryotes in the studied biofilms. Both microscopy-

based and metabarcoding approaches recovered strong relationships between community composition and impairment, which means that molecular characterization of biofilm eukaryotic communities can be an effective tool for monitoring stream health. The results of this study demonstrated good reproducibility of the metabarcoding using hyperviariable 18S-V9 region and indicated that several groups of molecularly characterized microbial eukaryotes had similarly robust response to impairment, although diatoms showed the strongest association with water quality. To further develop metabarcoding approach to stream monitoring, we suggest expanding this project by 1) sampling a wider range of wadeable streams in various ecoregions of New Jersey, 2) conducting a simultaneous assessment of water chemistry, habitat characteristics, macroinvertebrate and diatom morphology-based assessment and metabarcoding of biofilms, 3) exploring several molecular markers in addition to 18S-V9 region to target various groups of organisms that may have been insufficiently sampled by 18S-V9 sequencing.

Introduction

A.1.Use of microbial eukaryotes for water-quality monitoring in rivers and streams.

Biological monitoring of rivers and streams most frequently involves assessment of fish, benthic macroinvertebrate and algal assemblages. Among various microbial eukaryotic organisms found in biofilms, only algae, and especially diatoms are routinely used for bioassessment of stream health. The use of algae in bioassessments depends on morphological identification under light and electron microscopy and on the knowledge of their indicative properties (Stevenson 2014). Many algal taxa are known to be good indicators of specific environmental conditions, while the diversity of assemblages has been shown to decline in particularly stressful conditions. The drawbacks of these morphology-based methods include relatively high costs, the need for a high-level taxonomic expertise, inconsistencies in identification, and a lack of clear boundaries among morphologically defined taxa. Many algal taxa as well as aquatic fungi and heterotrophic protists have few morphological characters that could be observed with conventional microscopy and, therefore, are not used in biomonitoring. Even for diatoms that, unlike many other protists, have character-rich cell skeletons, identification to species level is often hampered by the presence of cryptic and pseudocryptic species (e.g., Vanormelingen et al. 2013), morphological plasticity, and unknown degree of geographic variability in species morphology. Ultimately, this leads to considerable uncertainties in environmental inferences and lower confidence of these biological indicators.

A.2. Molecular approach to characterization of microbial assemblages

DNA metabarcoding is a rapidly evolving methodology that allows for automated identification of organisms from environmental samples and is thus less affected by human bias (Ji et al. 2013, Pawlowski et al. 2018, Taberlet et al. 2012). This approach is based on high-throughput next-generation sequencing (NGS) of short DNA fragments from whole communities and the obtained sequences are then matched to reference sequences with known taxonomic assignments (Keck et al. 2017). Metabarcoding is known for its high sensitivity and ability to detect the presence of rare or low-abundance species (Zhan et al. 2013). It is quickly becoming less expensive and less time- and labor-consuming compared to traditional morphological

methods to identify taxa (Pawlowski et al. 2012, Bourlat et al. 2013). In the last few years, DNA metabarcoding has been successfully used to evaluate microbial communities, including eukaryotes, in various environments including oceanic (de Vargas et al. 2015; Grattepanche et al. 2016, Lindeque et al. 2013, Rachik et al. 2018, Xu et al. 2017) and lake (Filker et al. 2016, Wang et al. 2014) plankton, marine benthos (Bik et al. 2012, Chariton et al. 2010), sediment cores (Stoof-Leichsenring 2012, 2015) and streams (e.g., Apothéloz-Perret-Gentil et al. 2017, Volant et al. 2016).

High-throughput technology is quickly developing, and while the first environmental metabarcoding studies mostly used the Roche 454 pyrosequencing technology, the preferred sequencing platform for the last 3-4 years has been Illumina. Some researchers are also exploring the use of the Ion Torrent platform and Single molecule real time (SMRT) sequencing developed by Pacific Biosciences (PacBio). The scientific community is also intensively experimenting with various laboratory protocols and molecular markers that can either amplify DNA of a wide variety of taxa or be more selective in targeting specific groups of organisms.

Because diatoms are the microorganisms most commonly used for freshwater biomonitoring, several studies have investigated the potential use of DNA metabarcoding of diatoms for routine bioassessments (Kermarrec et al. 2013, 2014, Kelly et al. 2018, Rivera et al. 2018, Rimet et al. 2018, Visco et al. 2015, Zimmermann et al. 2011, 2015). A variety of markers have been explored on different sequencing platforms for barcoding diatoms (Vasselon et al. 2017, Apothéloz-Perret-Gentil et al. 2017, Zimmermann et al. 2011), such as the V4 hypervariable region of the 18S rDNA gene (Zimmermann et al. 2011, Luddington et al. 2012, Visco et al. 2015), D2/D3 region of the LSU rDNA gene (Hamsher et al. 2011), the ITS regions (Moniz & Kaczmarska 2009, 2010), and fragments of the rbcL (Hamsher et al. 2011, Macgillivary & Kaczmarska 2011) and cox1 genes (Evans et al. 2007, Evans & Mann 2009). Experiments conducted by Zimmermann et al. (2011) and Kermarrec et al. (2013) identified rbcL and 18S fragments as the most efficient molecular markers for discriminating diatom taxa at a level suitable for biomonitoring purposes. There is a growing consensus that the rbcL gene is the most suitable for metabarcoding of diatoms as it is a protein-coding gene, which allows for easily aligned sequences without significant intragenomic variability (Kelly et al. 2018). The main advantage of using 18S metabarcoding is its ability to amplify a wide variety of eukaryotic taxa, including diatoms. Curated and updatable libraries of reference diatom sequences (Rimet et

al. 2015, Zimmermann et al. 2014) were developed for these genes. The effects of laboratory procedures, such as methods of DNA extractions and PCR bias were investigated (Vasselon et al. 2016, 2018).

In addition to diatoms, other groups of microbial organisms have been explored as subjects of DNA metabarcoding assessments. An attractive option is an amplification of the same molecular marker across the whole eukaryotic community (Aguilar et al. 2014; Volant et al. 2016). Several authors have even attempted to characterize all prokaryotes and eukaryotes simultaneously in river biofilms (Mesa et al. 2017) or river water columns (Van Rossum et al. 2018). Separate amplifications of 16S and 18S gene fragments are usually performed simultaneously to estimate a more complete microbial community structure (e.g., Tragin et al. 2018). For example, amplification of prokaryotic 16S (Zankarini et al. 2017) or plastid-specific UPA (Sherwood et al. 2017) gene fragments allows for molecular identification of both bacteria and algae. Characterizing microbial communities in totality opens ample opportunities for unraveling interspecific interactions and complex ecological networks of organisms from specific habitats.

A.3. Objectives of the project

The goal of this pilot project was to evaluate the feasibility of a DNA metabarcoding approach for biological monitoring of New Jersey rivers and streams. Because multiple methods for NGS of environmental samples have been proposed and there is no consensus yet in which DNA marker or bioinformatics pipeline is the gold standard, we decided to try a method recommended by the Earth Microbiome Project (EMP, Thompson et al. 2017). We sequenced the V9 hypervariable region of 18S eukaryotic gene on the Illumina MiSeq platform. This gene is the most commonly sequenced in protists; therefore, there are many sequences available in GenBank that correspond to a vast variety of microbial eukaryotes. An obvious advantage of using this universal marker is characterization of the whole microbial eukaryote community as opposed to capturing only a portion of biodiversity if using markers specific to various taxonomic groups. Another advantage of using this particular marker was the availability of a complete protocol for constructing amplicon libraries using one-step PCR, which greatly reduces the cost of sequencing if when carried out on a large scale. The EMP protocol has previously

been used to characterize eukaryotic plankton communities of the open ocean (de Vargas et al. 2015) and estuaries (Abad et al. 2016).

The New Jersey DEP has sponsored several projects focused on river health assessment and abundant information is available on the water quality and biota in the NJ streams. Because most of New Jersey DEP's water quality biomonitoring methods were developed for wadeable streams, we chose to focus on these habitats for this pilot project. Our goal was to obtain biofilm samples from sites that cover a disturbance gradient from high-quality to severely degraded and explore their respective microbial eukaryotic community response to disturbance.

To evaluate possible advantages and disadvantages of a metabarcoding approach for stream health assessment compared to one based on morphological identification of organisms, we also carried out diatom enumeration in the same samples. In contrast to other taxonomic groups of protists, diatoms of New Jersey rivers and streams are well-studied (Ponader et al. 2007) and various region-specific diatom metrics are available for relating diatom assemblage composition to environmental stressors (e.g. Potapova et al 2004, Potapova & Charles 2007). Diatoms are also often a conspicuous component of river biofilms and, therefore, are well-suited for evaluating differences between molecular- and morphology-based bioassessments.

B. Methods

B.1. Sample collection

Fourteen sites throughout New Jersey were sampled in this study (Figure 1). Sites were selected based on historical water quality data available from the Water Quality Portal (https://www.waterqualitydata.us/), US Geological Survey Surface-Water Data for New Jersey (https://waterdata.usgs.gov/nj/nwis/sw) and the Delaware River Watershed Initiative (DRWI) (https://4states1source.org/), representing a gradient of high to low water quality in terms of nutrients, conductivity, and in-stream and riparian habitat conditions (Ponader et al. 2007). Three sites: Dunnfield Creek, Flat Brook and Neldons Brook were considered to be minimally impaired based on low nutrient and conductivity values (Table 1). Dunnfield Creek and Flat Brook have also been characterized earlier as high-quality sites based on macroinvertebrate IBI values (Kroll 2013). Four sites (Mulhockaway Creek, Musconetcong R. Lower, Musconetcong R. at Point

Mountain and Pequest R.) were considered to be moderately impaired based on water-quality data. These sites had Total Nitrogen (TN) values between 1 and 3 mg L^{-1}, Total Phosphorus (TP) between 0.02 and 0.04 mg L^{-1}, and conductivity between 336 and 645 μS cm^{-1}. Seven other sites (Musconetcong R. at Stephens State Park, Assunpink Creek, Rahway R. near Springfield, Rahway R. at Rahway, Passaic R., Ramapo R., and Saddle R.) were classified as heavily impaired based on comparatively high nutrient concentrations and/or high conductivity. These sites, except Musconetcong R. at Stephens State Park, had total phosphorus values at or higher than 0.1 mg L^{-1}, while Musconetcong R. at Stephens State Park site was classified as heavily impaired based on a relatively high conductivity value (708 μS cm^{-1}).

Samples of epilithic biofilms were collected between June and August 2017 (Table 1). At each site, three replicates of biofilm samples were collected, each from a single fully submerged rock using the 'top-rock' scraping method (Moulton et al. 2002) and suspended in ambient stream water in 250 mL bottles. Rocks were semi-quantitatively selected to represent in-stream habitat variability, where present (e.g. shallow vs. deep water; fast- vs. slow-moving water; less vs. greater canopy cover). Additionally, water temperature, pH, and conductivity were measured at each site (Table 1).

B.2. Laboratory methods – sample preparation

Fifty-milliliter aliquots of each sample were transferred to 50 mL conical-bottom tubes and centrifuged for 5 minutes. Supernatant stream water was aspirated using a vacuum pump. Each remaining pellet was resuspended with an additional aliquot of sample, up to 50 mL. Samples were re-centrifuged, aspirated, and re-suspended until all stream water was removed from each sample, leaving each with a single pellet of biomass.

B.3. Morphology-based diatom identification and enumeration

Biomass samples were re-suspended in deionized water and digested in 30% H_2O_2 to remove organic material. To remove acid, digested samples were centrifuged at 4,000 RPM for 10 minutes. Supernatant was decanted, and pellets were re-suspended in deionized water. This

rinsing process was repeated five times. Permanent slides of each subsample were prepared using Naphrax mounting medium (Brunel Microscopes, UK). The slides were examined with a Zeiss AxioImager A1 light microscope (LM) equipped with an AxioScopeMRm digital camera using DIC under oil immersion at 1000-1600x total magnification. At least four hundred valves were identified and enumerated to the lowest taxonomic level using standard floras (Hofmann et al. 2013, Krammer & Lange-Bertalot 1986, 1988, 1991a,b, Spaulding et al. 2018) and other current sources of diatom identification.

B.4. Amplicon library preparation and sequencing

Biofilm pellets were frozen at -20^0 C. Genomic DNA was extracted from each sample, in duplicate, using the Qiagen /MoBio PowerSoil DNA Isolation Kit following manufacturer's instructions. DNA concentration of the elutant was quantified using a QuBit 2.0 fluorometer prior to amplification, using Quibit dsDNA high sensitivity assay kits (Life Technologies, Thermo Fisher Scientific, Carlsbad, CA, USA).

NGS library preparation was carried out following a modified version of the Earth Microbiome Project (Gilbert et al. 2010) 18S Illumina Amplicon Protocol, version November 2016 (Caporaso et al. 2012). Forward and barcoded reverse primer constructs were those designed by L.W. Parfrey for the Illumina MiSeq platform, targeting microbial eukaryotes using the hypervariable v9 region of 18S SSU rRNA on the basis of primers developed by Amaral-Zettler et al. (2009). The forward primer construct consisted of 5' Illumina adapter, forward primer pad, forward primer linker and forward primer 1391f:
AATGATACGGCGACCACCGAGATCTACAC TATCGCCGTT CG
GTACACACCGCCCGTC

The reverse barcoded primer construct consisted of reverse complement of 3′ Illumina adapter, Golay barcode, reverse primer pad, reverse primer linker and reverse primer EukBr:

CAAGCAGAAGACGGCATACGAGAT XXXXXXXXXXXX AGTCAGTCAG CA
TGATCCTTCTGCAGGTTCACCTAC

PCR reaction mixtures were prepared using 1.0 μL template DNA, 13.0 μL PCR-grade water, 10.0 uL Platinum Hot Start PCR Master Mix, 0.5 μL of forward primer and 0.5 μL of barcoded reverse primer, each at a 10 μM concentration, for a total reaction volume of 25 μL. DNA

amplification was carried out using the following thermocycler conditions: an initial denaturing for 3 min. at 94 degrees C, 35 replication cycles of 94 degrees C for 45 seconds, 57 degrees C for 60 seconds, and 72 degrees C for 90 seconds. A final extension of 72 degrees C for 10 minutes was carried out. DNA extraction aliquots were not amplified in triplicate as recommended by the Earth Microbiome 18S protocol to reduce the risk of contamination during either the amplification or pooling steps. PCR products were visualized using gel electrophoresis to verify amplification and a target amplicon size of ~260 +/- 50 bp. PCR products were purified using a Qiagen QIAquick Purification kit following manufacturer's instructions and quantified using a QuBit fluorometer. Equal amplicon masses of each sample were pooled and submitted for paired-end (2 x 300 bp) sequencing on the Illumina MiSeq platform. The sequencing was carried out at the QB3 Vincent J. Coates Genomics Sequencing Lab, UC Berkley.

B.5. Bioinformatics pipeline

Sequence analysis was carried out floowing the R tutorial 'A DADA2 workflow for Big Data: Paired-end' (Callahan et al. 2016b). Demultiplexed paired-end sequences (2x300 base-pare reads) and primers and adapters were trimmed using Cutadapt v. 1.18 (Martin 2011). The resulting V9 fragments were on average 127 bp long. The reads were then checked for quality, denoised, truncated and merged using DADA2 workflow. Several truncation options in DADA2 were explored because the number of sequences obtained by this package heavily depend on the degree of trimming, and because the choice of truncation length is rather arbitrary and based on the visual inspection of sequence quality plots. After sequences were trimmed, our quality plots did not exhibit an obvious sharp decline in sequence quality, and therefore the reads could have been truncated at different lengths. The results reported in this report were obtained with truncating forward reads at 82 bp and reverse reads at 98 bp. When longer reads were merged, fewer sequences were produced, but the usual recommendation from the DADA2 developers was to truncate sequences at the shortest possible length because the merging algorithm tends to discard sequences with lower quality scores, thus leaving out too many rare sequences. Chimeras were identified and removed, and taxonomic assignment of sequences was done using the RDP Naïve Bayes Classifier algorithm (Wang et al. 2007) and aligning to the Silva reference database v. 128 (Pruesse et al. 2007), using subsets of OTUs clustered at 97% identity with consensus

taxonomy with >=80% bootstrap support. OTUs were identified to genus or lowest possible taxonomic level. For diatom OTUs identified at or above 'Class' level, GenBank BLAST searches were carried out to further verify their taxonomic identities.

For downstream analysis, we used the R package 'phyloseq' to combine the resulting taxonomy table with the respective matrix of OTUs by sample and associated field data (https://github.com/joey711/phyloseq). We used the analysis wrapper functions to filter OTUs to remove any non-eukaryote OTUs. We assigned OTUs to major taxonomic groups and constructed barplots to explore diversity and abundance within and among sites (Figs 2-5). We then subset our data by major eukaryote group and Hellinger-transformed each OTU matrix. We constructed non-constrained (NMDS) ordinations with fitted environmental variables to visualize sample-environment relationships for all data sets. For all data sets, we conducted PERMANOVA analyses to test for significant differences among sites in assemblage composition using a Bray-Curtis dissimilarity and 'Site' as the treatment. To test the effect of impairment on assemblage composition, we conducted Canonical Correspondence Analyses (CCA) with impairment category as a single constraining variable. We evaluated PERMANOVA model performance and CCA results to determine whether discrimination between site and impairment category was better predicted by all eukaryote OTUs, diatom OTUs only, or those of a single major eukaryote group.

Morphology-based diatom methods were initially assessed by calculating diatom-inferred nutrient metrics for each sample (Potapova and Charles 2007). We plotted TN and TP metric scores by site, arranged in average increasing order. We also constructed an NMDS ordination and a CCA constrained by impairment category for our diatom count data set. We used the same method as above to perform a PERMANOVA on transformed morphological assemblage data by site. We used the CCA and PERMANOVA to compare the performance of diatom metabarcoding vs. morphology-based methods.

C. Results and Discussion

C.1. DNA metabarcoding

C.1.1. Taxonomic assignment of OTUs

A total of 6,995,104 sequences obtained by the DADA2 procedure belonged to 5,869 OTUs (Appendix 1). 3,831 of these OTUs, representing approximately 37% of total abundance, were only assigned to the level of "Eukaryota" using the SILVA reference database and could not be assigned any lower-level taxonomic group. A subsequent BLAST search for these "unassigned eukaryotic" OTUs assigned 3,202 of these, leaving only 629 (11%) left as "unassigned Eukaryote" in this report. These unassigned OTUs represent only 1.7% of total abundance. The high proportion of unassigned Eukaryotes using the SILVA database may be due to the relatively low Eukaryote sequence diversity available in the reference database (104,020 18S sequences) and because the 18S-V9 region is found at the end of the 18S gene.

The most diverse (assigned) eukaryotic lineage in our dataset (2,717 OTUs) was the SAR supergroup that includes Stramenopiles, Alveolates and Rhizaria. Stramenopiles, also known as Heterokonts, constitute a major evolutionary lineage of heterotrophic and phototrophic eukaryotes characterized by the presence of two different flagella. This group is extremely diverse morphologically with representatives ranging from nanoflagellates to giant kelp species. Among Stramenopiles, the most diverse were diatoms (359 OTUs), Peronosporomycetes, or water moulds (159 OTUs) and Chrysophyceae, or golden algae (129 OTUs). Among other groups of Stramenopiles, the genus *Blastocystis*, which includes some representatives potentially pathogenic for humans (Ramírez et al. 2017), was detected at very low abundance. Alveolata that in freshwaters are mostly represented by heterotrophic ciliates and dinoflagellates were almost as diverse as Stramenoplies. Most Rhizaria OTUs belonged to Cercozoa.

The second most diverse major group in the whole dataset were Opisthokonta with 1775 unique OTUs. This lineage includes two major groups, fungi and animals (subgroup Metazoa within Holozoa). In our dataset, there were 1025 unique fungal OTUs and 750 unique Holozoan OTUs, mostly assigned to Arthropods (288 OTUs), Nematodes (95 OTUs), and Rotifers (85 OTUs).

The next major group in terms of OTUs diversity was Excavata with 247 unique OTUs. Although the Excavata sequences were not especially abundant and most of their representatives were common freshwater euglenoids, this group also included representatives of some genera (e.g. *Naegleria*) that include species known to be human pathogens (Siddiqui & Khan 2012). No pathogenic species from these genera were detected in this study, but the ability of the method to detect such sequences is notable.

In the Archaeplastida clade (499 OTUs), most OTUs (366) were classified as Chlorophyta, including green algae. 100 OTUs were identified as Charophyta, including land plants, while 33 OTUs were assigned to Rhodophyta, or red algae.

The less diverse eukaryotic groups were Amoebozoa (398 OTUs) and Cryptophyceae (63 OTUs). These groups were not abundant in terms of the number of sequences.

The most abundant OTUs across the dataset were sequences of several diatoms, such as *Melosira varians, Navicula* and *Cyclotella* spp., unspecified diatoms, bdelloid rotifers, unspecified Metazoans (animals), *Saprolegnia* sp. (water mould), and green alga *Stigeoclonium* sp. The prevalence of the diatom and holozoan sequences reflect the well-known dominance of diatoms and benthic invertebrates in the stream benthic communities.

C.1.2. Among- and within-site variation of taxonomic composition

Figure 2 shows the numbers of sequences grouped by major eukaryotic groups in 14 stream sites with six replicates (three field and two lab replicates pooled). The three major groups almost invariably were diatoms, holozoans and green algae, but the proportions of various groups somewhat varied. The largest difference was in between shaded sites (e.g., Neldons Bk.) that had much higher proportions of heterotrophic organisms and open sites that had larger populations of algae. As the sites in Figure 2 are ordered along the disturbance gradient, it is obvious that the proportions of major groups were unrelated to impairment.

Figures 3 and 4 demonstrate that among-site variability of taxonomic composition of the assemblages was greater than variability among field and lab replicates, thus indicating adequate quantitative assessment of the abundance of various taxa using metabarcoding and good reproducibility of the method.

Although the numbers of unique OTUs varied considerably among sites and, to a lesser extent, among field and lab replicates (lower plot in Figure 5), the proportions of major taxonomic groups were surprisingly stable across the samples (upper plot in Figure 5). Although too few other studies of total eukaryotic assemblages in epilithic biofilms are available for comparison, Zancarini et al (2017) also reported rather stable proportions of diatom, cyanobacterial and green algal OTUs across four river sampling sites. Levi et al. (2017) found rather similar taxonomic composition of eukaryotic assemblages in samples collected from different river sites but from the same type of substrate; the most variation in that study was found among substrates. In studies of marine eukaryotic plankton, proportions of major taxonomic groups were also found to be quite similar of the same size fractions in samples collected across large geographic areas (de Vargas et al. 2015)

The ordinations of all eukaryotic taxa in the dataset, either including all or excluding rare OTUs showed clear alignment of assemblages along a disturbance gradient. Moreover, as Figure 6 shows in an NMDS ordination of all eukaryotic OTUs, there was clear clustering of field and lab replicates representing different sites along a disturbance gradient. This indicates strong reproducibility of the results obtained by the method we used in this study, but other studies have reported low reproducibility and their insufficient accuracy for quantitative assessments (Leray & Knowlton 2017, Vasselon et al. 2018, Zimmermann et al. 2015). The reasons for these deficiencies include, but are not limited to PCR bias, or better amplification of certain taxa compared to others, highly variable copy numbers of marker genes in various taxa, the overriding effect of high-biomass organisms, and the influence of extraction methods (Vasselon et al. 2017). Although promising, DNA metabarcoding methods are not yet sufficiently tested to replace current methods for routine biomonitoring and further work is necessary to adapt them to that purpose.

C.2. Morphology-and molecular-based characterization of diatom assemblages

As a result of microscopical identification, a total of 236 diatom taxa were found in 42 examined samples (Appendix 2). The taxa with the highest number of occurrences were *Cocconeis placentula* (41), *Achnanthdium minutissimum* (40), *Amphora pediculus* (36), *Sellaphora atomus* (36), *Rhoicosphenia abbreviata* (32), *Achnanthidium rivulare* (31), *Navicula*

gregaria (30) and *Nitzschia amphibia* (30). The species that reached maximum abundance were *Achnanthidium rivulare (77%), Sellaphora atomus (75%), Cocconeis placentula (52%), Achnanthidium pyrenaicum (45%), Rhoicosphenia abbreviata (32%),* and *Hippodonta pseudacceptata (29%).* These taxa are commonly reported from rivers and streams of the eastern United States and New Jersey in particular (Potapova et al. 2004, Ponader et al. 2007).

A few diatom taxa found in this survey are notable because they have not been previously reported for the New Jersey rivers and streams. The first is *Navicula peregrinopsis* (Figure 7), that reached considerable abundance in all heavily impaired sites except Assunpink Cr. and was especially abundant in two sites with exceptionally high conductivity: Rahway R. near Springfield and Saddle R. Examination of slides housed at ANSP Diatom Herbarium and representing diatom samples collected in the same watercourses in 2001-2009 showed that in the past this species was present at very low abundance and either was not included in the counts at all or, later, when its abundance started to increase, was misidentified and reported as other species. *N. peregrinopsis* was only recently described from Holocene brackish-water coastal sediment in Connecticut (Witkowski et al. 2000) and has not been included in floras typically used for identification of river diatoms. It appears to be spreading and increasing in abundance because of the wide-spread salinization of urban streams in the recent decades.

Another rare species, *Ellerbeckia arenaria,* was found in samples from Musconetcong R. (M3), but not identified in the counts because of its low abundance. In fact, only frustule fragments of that species were encountered. This large-celled centric diatom is not usually found in streams and rivers but is more characteristic of subaerial habitats and large alkaline lakes. It is not clear at this point why this species appeared in Musconetcong R. Several sequences that are assigned the name of "*Paralia* (*Ellerbeckia*) *sol*" appear exclusively in M3 and two other sites from Musconetcong R. at relatively high abundance. It is clear that these sequences are in fact those of *Ellerebeckia arenaria*, but there was no overlap between our *E. arenaria* sequences as the only 18S partial sequence of this taxon in the GenBank does not reach the V9 region. This illustrated two points: first, that the metabarcoding is extremely sensitive because it detected this diatom in all sites along the Musconetcong R. and second, that it is heavily biased towards detection of large-celled organisms.

The latter point of overrepresentation of sequences of the large-celled species is also illustrated by the high abundance of the sequences of another large-celled diatom, *Melosira varians,* in our samples. This species was also frequently observed on slides, but not necessarily at a high abundance. The mismatch between the prevalence of *Melosira varians* sequences in the metabarcoding results and low to moderate cell numbers observed under the microscope was also noted by Kelly et al. (2018) in samples from British rivers. The cells of this diatom have high volume and apparently relatively high copy numbers of 18S rDNA. Several OTUs that matched the genus *Navicula,* but not assigned to any particular species were likewise very abundant. Morphological examination showed that *Navicula* species were quite diverse and abundant in our dataset, and some, especially *N. peregrinopsis*, had a very large cell volume, which may explain their relatively high contribution to OTUs abundance. The other abundant sequences were assigned to *Rhoicosphenia abbreviata, Gomphonema* sp., *Cocconeis* sp., *Eolimna minima (=small-celled Sellaphora* spp., *including S. atomus + S. nigri), Amphora* sp., *Ulnaria* sp., *Sellaphora* spp., and *Cyclotella meneghiniana*, all taxa common on the slides. A notable discrepancy was the relatively low abundance of sequences from small-celled diatoms from the family Achnanthidiaceae, which were often very abundant on slides. This may be explained by both low cell volume and insufficient coverage of these taxa in the reference database. Most of the 18S rDNA sequences of Achnanthidiaceae available in GenBank are too short and do not reach the V9 region of the gene.

In order to investigate the response of diatom assemblages to impairment, two metrics (TN-diatom index and TP –diatom index) developed by Potapova & Charles (2007) for diatom-based assessment of nutrient enrichment of rivers of eastern US were calculated from count data. As can be seen in Figure 8, the values of both metrics showed a nearly perfect relation to the degree of stream impairment. There was some variability among field replicates in moderately impacted sites, but this is expected if only a single rock is sampled. In most routine biomonitoring programs, multiple subsamples are pooled per sampled site and a composite sample is subsequently analyzed, thus smoothing out the within-site assemblage variability.

Unconstrained (NMDS) ordinations of diatom count data, displaying subsamples and species (Figures 9 & 10, respectively) showed that the diatom assemblages were highly responsive to impairment. Figure 10 displays taxa reaching 10% relative abundance in at least one sample,

though all taxa were included in the ordination. Taxa are labeled using corresponding "short names" in Appendix 2. The first NMDS axis reflects the major gradient in species composition, which corresponds to a disturbance gradient with minimally impacted sites (Dunnfield CR., Flat Bk. and Neldons Bk.) and corresponding dominant species positioned on the right and the most impacted sites on the left. This "disturbance gradient" is positively correlated to nutrient concentrations and conductivity. The second NMDS axis was mostly related to pH and separated diatom assemblages of Assunpink creek from the other sites. Unlike the other 13 sites, Assunpink creek has naturally soft water, which is still reflected by the composition of diatom assemblages despite of the masking effect of pollution. The NMDS also showed that within-site variation was lower than among-site variation in minimally impaired sites, but field replicates of the moderately and heavily impaired sites were often not clustered together, indicating a comparable degree of within- and among-site variability. This is likely caused by the homogenization of diatom flora in disturbed sites that are becoming more similar in their diatom species composition when experiencing similar impacts.

NMDS ordination of all diatom molecular OTUs also revealed a clear response to disturbance, with obviously tighter clustering of most site subsamples (Figure 11) as compared to the NMDS ordination of count data. Taxonomic assignment of OTUs that reached at least 20% relative abundance in at least one subsample (and shown in Figure 12) revealed similar trends to taxa identified by morphology. OTUs likely corresponding to morphologically identified taxa are plotted with red text, while OTUs not appearing in count data are plotted with green text (Figure 12), all using taxon "short names" (Appendix 1). Notably, DNA metabarcoding revealed a high abundance of OTUs corresponding to abundant taxa found in diatom counts, including *Cocconeis* spp., *Rhoicosphenia abbreviata, Melosira varians, Gomphonema* spp., *Ulnaria ulna, Amphora* spp., and small-celled naviculoid diatoms (e.g. *Eolimna minima, Sellaphora seminulum,* and *Mayamaea atomus*). DNA metabarcoding also identified abundant taxa known to inhabit low-pH and low-conductivity environments, including *Frustulia* spp., and taxa inhabiting moderately- to heavily impaired sites in terms of nutrient concentrations and conductivity, including *Amphora pediculus,* small-celled naviculoid diatoms, and *Pleurosira laevis.* Interestingly, *Achnanthidium minutissmum* was found in high abundance in low-conductivity sites using both morphological and molecular methods. As previously noted, the large diatoms are either moderately abundant (*Melosira varians*) or very rare (*Ellerbeckia*

arenaria) in counts, but found in high abundance using DNA metabarcoding. The congruence between count data and OTUs in this study not only indicates especially high reproducibility of diatom diversity using DNA metabarcoding, particularly in minimally impacted sites, but also a higher likelihood of capturing the entire diatom assemblage within a site.

NMDS ordinations of some other major eukaryotic groups showed significant response to environment as well. As seen in Figure 13, Chlorophyta (green algae) assemblages in minimally impaired sites showed clear separation from moderately and severely impaired sites, and within-site variability was clearly lower than among-site variability for the minimally impaired sites. Holozoa and "SAR_other" demonstrated a similar pattern of variation among impairment groups (Figures 14 and 15, respectively).

Among-site differences were significant for each data set, providing further support for high reproducibility of morphology-based and metabarcoding (Table 2). The response to impairment was significant in all data sets, as indicated by CCA ordinations constrained by impairment category only (Table 3). The F-ratios indicate that the molecular diatom data had the strongest response to impairment, while Chlorophyta (green algae) molecular data demonstrated the second strongest response, even stronger than that of the diatom microscopic count dataset.

Although our analyses show that molecularly characterized eukaryotes may serve as good environmental indicators, there are no metrics yet developed to measure water quality based on molecular data, as is the case for diatom counting methods. However, the metabarcoding method can more clearly distinguish among-site differences than traditional methods can, and OTUs positioned at both ends of the disturbance gradient are good candidates to be indicators of water quality.

The discrepancies between molecular and morphology-inferred community structure are often reported for protists (Elbrecht & Leese 2015), and are considered as the main deficiency of metabarcoding approach. One problem is a so-called PCR-bias, which is a tendency for the PCR procedure to amplify the DNA fragments with variable efficiency in different taxa (e.g., Kanagawa 2003). A related problem is the highly variable gene copy numbers among the taxa. For example, Godhe et al. (2008) demonstrated various diatom species differ in the number of 18S rDNA molecules per cell by several orders of magnitude, and the copy numbers are

positively correlated with cell biovolume. Since larger organisms have more DNA, their random occurrence in small subsamples used for DNA extraction may dramatically change the proportions of OTUs in the resulting data (e.g., Elbrecht et al. 2017).

C.3. Further steps towards developing metabarcoding of biofilm communities for the purposes of stream biomonitoring in New Jersey

The method we used in this pilot project (sequencing V9 fragment of 18S rDNA using the Earth Microbiome Protocol) was very successful in characterizing the diversity of microbial eukaryotes in stream biofilms. Their assemblages showed a strong and consistent response to stream impairment. Although insufficient reproducibility of DNA metabarcoding for stream bioassessment has been reported by other authors, the molecular data obtained in our study revealed lower among-replicate than among-site variation. DNA metabarcoding of diatoms outperformed morphological counts, showing a stronger relationship to environmental impairment. In minimally disturbed and heavily degraded sites, it appeared that DNA metabarcoding adequately captured the "true" diversity of diatom sequences, but not necessarily in moderately impacted sites. This is not surprising, as moderately impaired sites are usually the most diverse (Townsend et al. 1997). Because we used a modified amplicon library preparation protocol in which we only performed one DNA extraction per site, we likely failed to capture the full diversity of microbial eukaryotes. The Earth Microbiome Project offers a similar 16S rDNA amplicon library protocol, designed for prokaryotes, with the same pooling protocol. While there is a greater risk of contamination of prokaryotic samples, this is less of a concern with eukaryotic DNA; therefore, pooling several amplicon PCR products could better capture diversity and reduce the effects of PCR bias.

Our ability to identify diatom OTUs below the Class level was limited by our reference DNA database, which reflects the low diversity of diatom 18S-V9 sequence fragments available in GenBank. While a wide diversity of partial 18S rDNA sequences are available, many do not include the V9 region, which is found at the end of the 18S gene. Therefore, we were not able to compare most taxonomic identities of our diatom sequences to those found in our counts. This finding underscores the need to obtain reference sequences for a large number of taxa commonly found in New Jersey streams.

The rbcL gene allows for finer taxonomic assignment of diatoms and therefore is worth considering for NGS-based bioassessment studies. A diatom-specific DNA reference database for rbcL, *R-Syst::diatom*, is rapidly developing in Europe for the purpose of environmental metabarcoding as it is gaining popularity as the preferred diatom-based bioassessment method for its reproducibility and specificity. For example, a recent study by Kelly et al. (2018) demonstrated a relatively reliable recovery of diatom community structure when using a rbcL gene fragment for sequencing diatoms. Unlike the Earth Microbiome Project procedure, which uses one-step PCR, the methods used by Kelly et al. (2018) and other authors who applied metabarcoding to stream diatom assemblages (Apothéloz-Perret-Gentil et al. 2017, Zimmermann et al. 2015, Vasselon et al. 2017) are based on dual PCR (Bourlat et al. 2016). No taxon-specific primer constructs suitable for the single-step PCR metabarcoding have been developed so far. In comparison to the EMP protocol, the two-step PCR considerably increases the cost of the lab portion of the metabarcoding and is another source of potential error. On the other hand, the two-step PCR procedures more often have been reported to produce consistent results as assessed by the mock-community experiments and replicate sampling (Kermarreck et al. 2013, Kelly et al. 2018). An obvious disadvantage of using taxon-selective primers is targeting only a single group of organisms instead of covering a vast variety of microbial eukaryotes. Additionally, developing metabarcoding methods for covering all eukaryotes and suitable for quantitative assessment is an area of intensive research.

Given the promising results of eukaryotic DNA metabarcoding in this pilot study, we suggest expanding this investigation by sampling a larger set of sites and testing reliability and reproducibility of the method by assessing additional sources of variability (seasonal, among-substrates, etc). We advise against modifying the existing EMP amplicon library preparation protocol and suggest pooling PCR products in triplicate to more adequately capture eukaryotic diversity and increase reproducibility among replicates. In order to ascertain taxonomic identities of eukaryotic groups poorly represented in existing 18S reference databases, we suggest further trials of metabarcoding covering a narrower range of taxa using the two-step PCR procedures. The following markers covering exclusively diatoms can be explored: rbcL (Kelly et al. 2018, Kermarrec et al. 2013, 2014, Stoof-Leichsenring 2012, 2015) and V4-18S (Zimmermann 2011). The UPA primers targeting a fragment of 23S plastid gene and thus amplifying most algal groups (Sherwood et al. 2017) can also be explored as a potential biomonitoring tool.

Although groups of eukaryotes other than diatoms can serve as valuable bioindicators, an advantage of using diatoms for these early steps of developing metabarcoding protocol is the ability for direct comparison of sequencing results to microscopic observations. An obstacle for such a direct comparison is, however, the lack of reference sequences in the public databases. Although initial reference databases have been developed for European taxa (Rimet at al. 2015), and many of those should be applicable to North American diatoms, many taxa in eastern US rivers are endemic to this region or are at least much more common than in the other parts of the world. Therefore, it is necessary to obtain sequences for many local taxa, which requires culturing and Sanger-sequencing of several loci that potentially could be used for metabarcoding. This labor-intensive process also requires a good knowledge of local diatom flora and the ability to distinguish unique species of living cells. As it was emphasized by several authors (e.g. Stein et al. 2014, Visco et al. 2015, Pawlowski et al. 2018), a considerable initial investment is required for the development of the metabarcoding bioassesment and its incorporation into routine biomonitoring programs, but once developed, this method will be much more cost-effective than morphology-based identifications.

All environmental metabarcoding studies show at least some discrepancies between molecular and morphology-based identification results. Therefore, it is important to include parallel molecular and morphological assessments in future studies for calibration purposes, as both approaches have their strengths and weaknesses and provide complementary information (Groendahl et al. 2017, Rimet et al. 2018).

D. Conclusions

Metabarcoding of microeukaryotic stream biofilm assemblages using the 18S-V9 Earth Microbiome Protocol and Illumina MiSeq technology produced repeatable results and revealed a vast diversity of organisms typically found in fresh waters, including a few potentially pathogenic protists, albeit detected at very low abundance. We found good correspondence between identities of diatoms observed under the microscope and revealed by next-generation amplicon sequencing.

While traditional microscopic identification of diatoms produced a reliable and solid assessment of stream impairment in this project, molecular signatures of diatoms had even stronger response to impairment. Metabarcoding thus represents a powerful tool to scale-up biomonitoring programs as it allows a rapid processing of many samples in a cost- and time-effective manner. At the same time, metabarcoding requires additional testing to estimate its reliability and potential sources of error. Reference sequence databases need further development in order to identify organisms present in study sites, especially for universal eukaryotic gene markers such as the V9 region of 18S. Calibrating metabarcoding methods against microscopy-based enumeration will be required at the stage of method development, if added to standard bioassessment protocols. An added benefit of microscopy is the potential to compare results of recent and past observations as the permanent diatom slides are easily available for examination in the public collections. For instance, we could track a recent expansion of a salt-tolerant diatom species in New Jersey streams, which would likely indicate a response to salinization. We envision that future biomonitoring programs will combine large-scale molecular assessments combined with occasional selective microscopical examination of the assemblages as the two approaches are likely to yield complementary results.

E. Acknowledgements

Dr. Chris Sales and Dr. Jacob Price, the Department of Civil, Architectural, and Environmental Engineering, Drexel University provided invaluable help by guiding us through all steps of the next-generation sequencing and molecular data analyses.

F. References

Abad, D., A. Albaina, M. Aguirre, A. Laza-Martínez, I. Uriarte, A. Iriarte, et al. 2016. Is metabarcoding suitable for estuarine plankton monitoring? A comparative study with microscopy. Marine Biology 163(7):1–13.

Abdul Majid, M.A., T. Mahboob, B.G.J. Mong, N. Jaturas, R.L. Richard, T. Tian-Chye & V. Nissapatorn. 2017. Pathogenic waterborne free-living amoebae: An update from selected Southeast Asian countries. PLoS ONE 12(2): e0169448.

Aguilar, M., E. Glücksman, D. Bass & J.B. Dacks. 2014. Next Generation Sequencing of Protists as a Measure of Microbial Community in Oil Sands Tailings Ponds: Amplicon Versus Metagenomic Approaches. Oil Sands Research and Information Network, University of Alberta, School of Energy and the Environment, Edmonton, Alberta. OSRIN Report No. TR-56. 24 pp.

Amaral-Zettler, L.A., E.A. McCliment, H.W. Ducklow & S.M. Huse. 2009. A method for studying protistan diversity using massively parallel sequencing of V9 hypervariable regions of small-subunit ribosomal RNA genes. PLoS ONE 4: e6372.

Amend, A.S., K.A. Seifert & T. Bruns. 2010. Quantifying microbial communities with 454 pyrosequencing: does read abundance count? Molecular Ecology 19: 5555–5565.

Apothéloz-Perret-Gentil, L., A. Cordonier, F. Straub, J. Iseli, P. Esling & J. Pawlowski. 2017. Taxonomy-free molecular diatom index for high-throughput eDNA biomonitoring. Molecular Ecology Resources 17: 1231-1242.

Bik, H.M., W. Sung, P. De Ley, J.G. Baldwin, J. Sharma, A. Rocha-Olivares & W.K. Thomas. 2012. Metagenetic community analysis of microbial eukaryotes illuminates biogeographic patterns in deep-sea and shallow water sediments. Molecular Ecology 21(5): 1048–1059

Bourlat, S.J., A. Borja, J. Gilbert, M.I. Taylor, N. Davies, S.B. Weisberg, J.F. Griffith, T. Lettieri, D. Field, J. Benzie, F.O. Glöckner. N. Rodríguez-Ezpeleta, D.P. Faith, T.P. Bean & M. Obst. 2013. Genomics in marine monitoring: new opportunities for assessing marine health status. Marine Pollution Bulletin 74(1): 19-31.

Bourlat, S.J., Q. Haenel, J. Finnman & M. Leray. 2016. Preparation of Amplicon libraries for metabarcoding of marine eukaryotes using Illumina MiSeq: The Dual-PCR Method. Methods in Molecular Biol.ogy 1452:197-207.

Callahan, B.J., P.J. McMurdie, M.J. Rosen, A.W. Han, A.J. Johnson & S.P. Holmes. 2016a. DADA2: high-resolution sample inference from Illumina amplicon data. Nature Methods 13: 581–583.

Callahan, B.J., K. Sankaran, J.A. Fukuyama, P.J. McMurdie & S.P. Holmes. 2016b. Bioconductor workflow for microbiome data analysis: from raw reads to community analyses. F1000Research 5: 1492.

Caporaso, J.G., J. Kuczynski, J. Stombaugh, K. Bittinger, F.D. Bushman, E.K. Costello, N. Fierer, A.G. Pena, J.K. Goodrich, J.I. Gordon, G.A. Huttley, S.T. Kelley, D. Knights, J.E. Koenig, R.E. Ley, C.A. Lozupone, D. McDonald, B.D. Muegge, M. Pirrung, J. Reeder, J.R. Sevinsky, P.J. Turnbaugh, W.A. Walters, J. Widmann, T. Yatsunenko, J. Zaneveld & R. Knight. 2010. QIIME allows analysis of highthroughput community sequencing data. Nature Methods 7: 335–336.

Caporaso, J.G., C.L. Lauber, W.A. Walters, D. Berg-Lyons, J. Huntley, N. Fierer & R. Knight. 2012. Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. The ISME Journal 6(8): 1621–1624.

Chariton, A. A., L.N. Court, D.M. Hartley, M.J. Colloff & C.M. Hardy. 2010. Ecological assessment of estuarine sediments by pyrosequencing eukaryotic ribosomal DNA. Frontiers in Ecology and the Environment 8: 233–238.

De Vargas, C.S., A.N. Audic, J. Decelle, F. Mahé, R. Logares, E. Lara, C. Berney, N. Le Bescot, I. Probert, M. Carmichael, J. Poulain, S. Romac, S. Colin, J.M. Aury, L. Bittner, S. Chaffron, M. Dunthorn, S. Engelen, O. Flegontova, L. Guidi, A. Horák, O. Jaillon, G. Lima-Mendez, J. Lukeš, S. Malviya, R. Morard, M. Mulot, E. Scalco, R. Siano, F. Vincent, A. Zingone, C. Dimier, M. Picheral, S. Searson, S. Kandels-Lewis, S.G. Acinas, P. Bork, C. Bowler, G. Gorsky, et al. Tara Oceans Coordinators. 2015. Eukaryotic plankton diversity in the sunlit ocean. Science 348: 1261605*a*

Elbrecht, V. & Leese, F. 2015. Can DNA-based ecosystem assessments quantify species abundance? Testing primer bias and biomass - sequence relationships with an innovative metabarcoding protocol. PLoS ONE 10(7): e0130324.

Elbrecht, V., B. Peinert & F. Leese. 2017. Sorting things out: Assessing effects of unequal specimen biomass on DNA metabarcoding. Ecology and Evolution 7: 6918–6926.

Emilson, C.E., D.G. Thompson, L.A. Venier, T.M. Porter, T. Swystun, D. Chartrand, S. Capell & M. Hajibabaei. 2017. DNA metabarcoding and morphological macroinvertebrate metrics reveal the same changes in boreal watersheds across an environmental gradient. Scientific Reports 7: 12777.

Evans, K. M. & Mann, D. G. 2009. A proposed protocol for nomenclaturally effective DNA barcoding of microalgae. Phycologia 48 (1): 70−74.

Evans, K. M. A.H. Wortley & D.G. Mann. 2007. An assessment of potential diatom “barcode” genes (cox1, rbcL, 18S and ITS rDNA) and their effectiveness in determining relationships in Sellaphora (Bacillariophyta). Protist 158: 349−364.

Filker, S., R. Sommaruga, I. Vila & T. Stoeck. 2016. Microbial eukaryote plankton communities of high-mountain lakes from three continents exhibit strong biogeographic patterns. Molecular Ecology 25: 2286–2301.

Gilbert, J.A., F. Meyer, J. Jansson, J. Gordon, N. Pace, J. Tiedje, et al. 2010. The Earth Microbiome Project: meeting report of the "1 EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6 2010. Standards in Genomic Sciences 3: 249–253.

Godhe, A., M.E. Asplund, K. Härnström, V. Saravanan, A. Tyagi & I. Karunasagar. 2008. Quantification of diatom and dinoflagellate biomasses in coastal marine seawater samples by real-time PCR. Applied Environmental Microbiology 74: 7174–7182.

Gong, J., J. Dong, X. Liu & R. Massana. 2013. Extremely high copy numbers and polymorphisms of the rDNA operon estimated from single cell analysis of oligotrich and peritrich ciliates. Protist 164: 369–379.

Grattepanche, J., L.F. Santoferrara, G.B. McManus & L. Katz. 2016. Unexpected biodiversity of ciliates in marine samples from below the photic zone. Molecular Ecology 25: 3987–4000.

Groendahl, S., M. Kahlert & P. Fink. 2017. The best of both worlds: A combined approach for analyzing microalgal diversity via metabarcoding and morphology-based methods. PLoS ONE 12(2): e0172808.

Hamsher, S. E., K.M. Evans, D.G. Mann, A. Poulickova & G.W. Saunders. 2011. Barcoding diatoms: exploring alternatives to COI-5P. Protist 162: 405−422.

Hardge, K., S. Neuhaus, E.S. Kilias, C. Wolf, K. Metfies & S. Frickenhaus. 2018. Impact of sequence processing and taxonomic classification approaches on eukaryotic community structure from environmental samples with emphasis on diatoms. Molecular Ecology Resources 18: 204–216.

Hofmann, G., H. Lange-Bertalot & M. Werum. 2013. Diatomeen im Süßwasser-Benthos von Mitteleuropa: 2 Corrected Edition Koeltz Scientific Books, Königstein, 908 pp.

Ji, Y., L. Ashton, S.M. Pedley, D.P. Edwards, Y. Tang, A. Nakamura, et al. 2013. Reliable, verifiable and efficient monitoring of biodiversity via metabarcoding. Ecology Letters 16(10): 1245–1257.

Keck, F., V. Vasselon, K. Tapolczai, F. Rimet & A. Bouchez. 2017. Freshwater biomonitoring in the information age. Frontiers in Ecology and Environment 15(5): 266–274.

Kelly, M., N. Boonham, S. Juggins, P. Kille, D. Mann, D. Pass, M. Sapp, S. Sato & R. Glover. 2018. A DNA based diatom metabarcoding approach for Water Framework Directive classification of rivers. Environmental Agency Report SC140024/R available at https://assets.publishing.service.gov.uk/government/uploads/system/uploads/attachment_data/file/684493/A_DNA_based_metabarcoding_approach_to_assess_diatom_communities_in_rivers_-_report.pdf

Kermarrec, L., A. Franc, F. Rimet, P. Chaumeil, J.M. Frigerio, J.F. Humbert & A. Bouchez. 2014. A next-generation sequencing approach to river biomonitoring using benthic diatoms. Freshwater Science 33: 349–363.

Kermarrec, L., A. Franc, F. Rimet, P. Chaumeil, J.F. Humbert & A. Bouchez. 2013. Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Molecular Ecology Resources 1: 607–619.

Krammer, K. & Lange-Bertalot, H. 1986. Bacillariophyceae. 1. Teil: Naviculaceae In: Ettl, H., J. Gerloff, H. Heynig and D. Mollenhauer (eds.) Süsswasserflora von Mitteleuropa, Band 2/1. Gustav Fisher Verlag, Jena. 876 pp.

Krammer, K. & Lange-Bertalot, H. 1988. Bacillariophyceae. 2. Teil: Bacillariaceae, Epithemiaceae, Surirellaceae. in Ettl, H., Gerloff, J., Heynig, H. and Mollenhauer, D. (eds) Süsswasserflora von Mitteleuropa, Band 2/2. VEB Gustav Fischer Verlag: Jena. 596 pp.

Krammer, K. & Lange-Bertalot, H. 1991a. Bacillariophyceae. 3. Teil: Centrales, Fragilariaceae, Eunotiaceae. in Ettl, H., Gerloff, J., Heynig, H. and Mollenhauer, D. (eds) Süsswasserflora von Mitteleuropa, Band 2/3. Gustav Fischer Verlag: Stuttgart, Jena. 576 pp.

Krammer, K. & Lange-Bertalot, H. 1991b. Bacillariophyceae. 4. Teil: Achnanthaceae, Kritische Ergänzungen zu Navicula (Lineolatae) und Gomphonema, Gesamtliteraturverzeichnis Teil 1-4. in Ettl, H., Gärtner, G., Gerloff, J., Heynig, H. and Mollenhauer, D. (eds) Süsswasserflora von Mitteleuropa, Band 2/4. Gustav Fischer Verlag: Stuttgart, Jena. 437 pp.

Kroll, S.A. 2013. Delaware River Watershed Initiative: Subwatershed Cluster Characterization Report. The Academy of Natural Sciences of Drexel University.

Levi, P.S., P. Starnawski, B. Poulsen, A. Baattrup-Pedersen, A. Schramm & T. Riis. 2017. Microbial community diversity and composition varies with habitat characteristics and biofilm function in macrophyte-rich streams. Oikos 126: 398-409.

Leray, M. & Knowlton, N. 2017. Random sampling causes the low reproducibility of rare eukaryotic OTUs in Illumina COI metabarcoding. PeerJ. 2017 Mar 22:5: e3006. doi: 10.7717/peerj.3006.

Lindeque, P.K., H.E. Parry, R.A. Harmer, P.J. Somerfield & A. Atkinson. 2013. Next Generation Sequencing reveals the hidden diversity of zooplankton assemblages. PLoS ONE 8(11): e81327.

Luddington, I.A., I. Kaczmarska & C. Lovejoy. 2012. Distance and character-based evaluation of the V4 region of the 18S rRNA gene for the identification of diatoms (Bacillariophyceae). PLoS ONE, 7: e45664.

Macgillivary, M.L. & Kaczmarska, I. 2011. Survey of the efficacy of a short fragment of the rbcL gene as a supplemental DNA barcode for diatoms. Journal of Eukaryotic Microbiology 58: 529−536.

Marciano-Cabral, F. & Cabral, G. 2003. *Acanthamoeba* spp. as agents of disease in humans. Clinical Microbiology Reviews 16(2): 273–307.

Martin, M. 2011. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal 17(10): 10-12.

Mesa, V., J.L.R. Gallego, R. González-Gil, B. Lauga, J. Sánchez, C. Méndez-García & A.I. Peláez. 2017. Bacterial, archaeal, and eukaryotic diversity across distinct microhabitats in an acid mine drainage. Frontiers in Microbiology 8:1756.

Moniz, M.B.J. & Kaczmarska, I. 2009. Barcoding micro- and mesofauna. Barcoding diatoms: is there a good marker? Molecular Ecology Resources 9: 65−74.

Moniz, M.B.J. & Kaczmarska, I. 2010. Barcoding of diatoms: nuclear encoded ITS revisited. Protist 161: 7−34.

Moulton, S.R., J.G. Kennen, R.M. Goldstein & J.A. Hambrook. 2002. Revised Protocols for Sampling Algal, Invertebrate and Fish Communities as Part of the National Water Quality Assessment Program. Open -File Report 02-150. U.S. Geological Survey, Reston, VA. 75 pp.

Pawlowski, J., S. Audic, S. Adl, D. Bass, L. Belbahri, C. Berney, et al. 2012. CBOL Protist Working Group: barcoding eukaryotic richness beyond the Animal, Plant, and Fungal kingdoms. PLoS Biology 10(11): e1001419.

Pawlowski, J., M. Kelly-Quinn, F. Altermatt, L. Apothéloz-Perret-Gentil, P. Beja, A. Boggero, A. Borja, A. Bouchez, T. Cordier, I. Domaizon, et al. 2018. The future of biotic indices in the ecogenomic era: Integrating (e)DNA metabarcoding in biological assessment of aquatic ecosystems. Science of The Total Environment 637–638: 1295-1310.

Ponader, K.C., D.F. Charles & T.J. Belton. 2007 Diatom-based TP and TN inference models and indices for monitoring nutrient enrichment of New Jersey streams. Ecological Indicators 7: 79–93.

Potapova, M., D.F. Charles, K.C. Ponader & D.M. Winter. 2004. Quantifying species indicator values for trophic diatom indices: a comparison of approaches. Hydrobiologia 517: 25–41.

Potapova, M. & Charles, D.F. 2007. Diatom metrics for monitoring eutrophication in rivers of the United States. Ecological Indicators 7: 48–70.

Pruesse, E., C. Quast, K. Knittel, B.M. Fuchs, W. Ludwig, J. Peplies & F.O. Glöckner. 2007. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic Acids Research 35: 7188–7196.

Rachik, S., U. Christaki, L.L. Li, S. Genitsaris, E. Breton & S. Monchy. 2018. Diversity and potential lactivity patterns of planktonic eukaryotic microbes in a mesoeutrophic coastal area (eastern English Channel). PLoSONE 13(5): e0196987.

Ramírez, J.D. A. Flórez, M. Olivera, M. Consuelo Bernal & J. Giraldo. 2017. *Blastocystis* subtyping and its association with intestinal parasites in children from different geographical regions of Colombia. PLoS ONE. 12. e0172586.

Rimet, F., P. Chaumeil, F. Keck, L. Kermarrec, V. Vasselon, M. Kahlert, A. Franc & A. Bouchez. 2015. R-Syst::diatom: a barcode database for diatoms and freshwater biomonitoring - data sources and curation procedure. INRA Report, 14 p.

Rimet, F., V. Vasselon, B. Keszte & A. Bouchez. 2018. Do we similarly assess diversity with microscopy and high-throughput sequencing? Case of microalgae in lakes. Organisms Diversity and Evolution 18:51.

Rivera, S., V. Vasselon, S. Jacquet, D. Ariztegui, A. Bouchez & F. Rimet. 2018. Metabarcoding of lake benthic diatoms: from structure assemblages to ecological assessment. Hydrobiologia 807: 37-51.

Sherwood, A.R., M.N. Dittbern, E.T. Johnston & K.Y. Conklin. 2017. A metabarcoding comparison of windward and leeward airborne algal diversity across the Koʻolau Mountain Range on the island of Oʻahu, Hawaiʻi. Journal of Phycology 53: 437-445.

Siddiqui, R. & Khan, N.A. 2012. Biology and pathogenesis of *Acanthamoeba*. Parasites & Vectors 5: 6.

Smith, K.F., G.S. Kohli, S.A. Murray & L.L. Rhodes. 2017. Assessment of the metabarcoding approach for community analysis of benthic-epiphytic dinoflagellates using mock communities. New Zealand Journal of Marine and Freshwater Research, 51(4): 555–576.

Spaulding, S.A., I.W. Bishop, M.B. Edlund, S. Lee & M. Potapova. 2018. Diatoms of North America. https://diatoms.org/.

Stein, E.D., B.P. White, R.D. Mazor, J.K. Jackson, J.M. Battle, P.E. Miller, E.M. Pilgrim & B.W. Sweeney. 2014. Does DNA barcoding improve performance of traditional stream bioassessment metrics? Freshwater Science 33(1): 302-311.

Stevenson, J. 2014. Ecological assessments with algae: a review and synthesis. Journal of Phycology 50(3): 437-461.

Stoof-Leichsenring, K.R., L.S. Epp, M.H. Trauth & R. Tiedemann. 2012. Hidden diversity in diatoms of Kenyan Lake Naivasha: a genetic approach detects temporal variation. Molecular Ecology 21: 1918–1930.

Stoof-Leichsenring, K., U. Herzschuh, L. Pestryakova, J. Klemm, L. Epp & R. Tiedemann. 2015. Genetic data from algae sedimentary DNA reflect the influence of environment over geography. Scientific Reports 5: 1-11.

Taberlet, P., E. Coissac, F. Pompanon, C. Brochmann & E. Willersley. 2012. Towards next-generation biodiversity assessment using DNA metabarcoding. Molecular Ecology 21(8): 2045-50.

Thompson, L.R., J.G. Sanders, D. McDonald, A. Amir, J.K. Jansson, J.A. Gilbert, R. Knight & The Earth Microbiome Project Consortium. 2017. A communal catalogue reveals Earth's multiscale microbial diversity. Nature 551: 457-463.

Townsend C.R., Scarsbrook M.R., Dolédec S. 1997. The intermediate disturbance hypothesis, refugia, and biodiversity in streams. Limnology and Oceanography 42(5): 938-949.

Tragin, M., A. Zingone & D. Vaulot. 2018. Comparison of coastal phytoplankton composition estimated from the V4 and V9 regions of the 18S rRNA gene with a focus on photosynthetic groups and especially Chlorophyta. Environmental Microbiology 20: 506-520.

Vanormelingen, P., K.M. Evans, V.A. Chepurnov, W. Vyverman & D.G. Mann. 2013. Molecular species discovery in the diatom *Sellaphora* and its congruence with mating trials. Fottea 13(2): 133–148.

Van Rossum, T., M.I. Uyaguari-Diaz, M. Vlok, M.A. Peabody, A. Tian, K.I. Cronin, M. Chan, M.A. Croxen, W.W.L. Hsiao, J. Isaac-Renton, P.K.C. Tang, N.A. Prystajecky, C.A. Suttle,

F.S.L. Brinkman. 2018. Spatiotemporal dynamics of river viruses, bacteria and microeukaryotes. BioRxiv preprint available at http://dx.doi.org/10.1101/259861.

Vasselon, V., A. Bouchez, F. Rimet, S. Jacquet, R. Trobajo, M. Corniquel, K. Tapolczai & I. Domaizon. 2018. A correction factor inferred from cell biovolume improves quantification in diatom metabarcoding. Methods in Ecology and Evolution 9:1060–1069.

Vasselon, V., I. Domaizon, F. Rimet, M. Kahlert & A. Bouchez. 2016. Application of high-throughput sequencing (HTS) metabarcoding to diatom biomonitoring: do DNA extraction methods matter? Freshwater Science, 36(1): 162–177.

Vasselon, V., F. Rimet, K. Tapolczai & A. Bouchez. 2017. Assessing ecological status with diatoms DNA metabarcoding: scaling-up on a WFD monitoring network (Mayotte island, France). Ecological Indicators 82:1 –12.

Visco, J.A., L. Apothéloz-Perret-Gentil, A. Cordonier, P. Esling, L. Pillet & J. Pawlowski. 2015. Environmental monitoring: inferring the Diatom Index from Next-Generation Sequencing data. Environmental Science & Technology 49 (13): 7597-7605.

Volant, A., M. Héry, A. Desoeuvre, C. Casiot, G. Morin, P.N. Bertin, et al. 2016. Spatial distribution of eukaryotic communities using high-throughput sequencing along a pollution gradient in the arsenic-rich creek sediments of carnoulès mine, France. *Microial Ecology* 72 608–620.

Wang, J., F. Wang, L. Chu, H. Wang, Z. Zhong, et al. 2014. High genetic diversity and novelty in eukaryotic plankton assemblages inhabiting saline lakes in the Qaidam Basin. PLoS ONE 9(11): e112812.

Wang, Q., G.M. Garrity, J.M. Tiedje & J.R. Cole. 2007. Naïve Bayesian classifier for rapid assignment of rRNA sequences into the new bacterial taxonomy. Applied and Environmental Microbiology 73(16): 5261-5267.

Witkowski, A., H. Lange-Bertalot & D. Metzeltin. 2000. Diatom Flora of Marine Coasts I Iconographia Diatomologica 7: 1-925.

Xu D., R. Li, C. Hu, P. Sun, N. Jiao & A. Warren. 2017. Microbial Eukaryote Diversity and Activity in the Water Column of the South China Sea Based on DNA and RNA High Throughput Sequencing. Frontiers in Microbiology 8: 1121.

Zancarini, A., I. Echenique-Subiabre, D. Debroas, N. Taïb, C. Quiblier & J. Humbert. 2017. Deciphering biodiversity and interactions between bacteria and microeukaryotes within epilithic biofilms from the Loue River, France. Scientific Reports 7: 4344.

Zhan, A., M. Hulak, F. Sylvester, X. Huang, A.A. Adebayo, C.L. Abbott, S.J. Adamowicz, D.D. Heath, et al. 2013. High sensitivity of 454 pyrosequencing for detection of rare species in aquatic communities. Methods in Ecology and Evolution 4: 558–565.

Zimmermann, J., N. Abarca, N. Enke, O. Skibbe, W.H. Kusber & R. Jahn. 2014. Taxonomic reference libraries for environmental barcoding: a best practice example from diatom research. PLoS ONE, 9, e108793.

Zimmermann, J., G. Glöckner, R. Jahn, N. Enke & B. Gemeinholzer. 2015. Metabarcoding vs. morphological identification to assess diatom diversity in environmental studies. Molecular Ecology Resources, 15: 526 –542.

Zimmermann, J., R. Jahn & B. Gemeinholzer. 2011. Barcoding diatoms: evaluation of the V4 subregion on the 18S rRNA gene, including new primers and protocols. Organisms Diversity and Evolution 11: 173.

G. Tables

Table 1. Locations and water-quality characteristics of sampling sites.

SITE NAME	Code	Latitude	Longitude	USGS or DRWI site code	Nearest AMNET Site	DATE	EC, µS/cm	pH	TN, mg/L	TP, mg/L	Impairment
Dunnfield Creek	D	40.97106	-75.12669	USGS 01442760	AN0012	3-Jun-17	39	6.8	0.1	0.01	Good
Flat Brook	F	41.10888	-74.95195	USGS 01440000	AN007-8	3-Jun-17	238	7.2	0.2	0.01	Good
Neldons Brook	N	41.08489	-74.82656	USGS 01443466	AN0023A	3-Jun-17	209	7.2	0.3	0.01	Good
Musconetcong R. Lower	M2	40.70397	-74.98880	DRWI NHML2	AN0072	21-Jun-17	610	8.4	2.7	0.02	Fair
Musconetcong at Point Mountain	M3	40.76749	-74.91166	DRWI NUM25	AN0069	21-Jun-17	640	8.4	1.9	0.03	Fair
Mulhockaway Creek	Mh	40.64750	-74.96888	USGS 01396660	AN0321	21-Jun-17	336	8.3	1	0.03	Fair
Pequest R.	P4	40.83719	-74.95371	DRWI NPQ28	AN0043	21-Jun-17	645	8.9	1.4	0.04	Fair
Musconetcong at Stephens SP	M5	40.87370	-74.80565	DRWI NLM26	AN0064	21-Jun-17	708	8.8	1	0.03	Poor
Assunpink Creek	A	40.21722	-74.76861	USGS 01464020	AN0116	1-Aug-17	494	7.2	5	0.5	Poor
Passaic River	P9	40.88472	-74.22611	USGS 01389500	AN0274	1-Aug-17	595	8.1	2.9	0.2	Poor
Rahway River near Springfield	R7	40.68750	-74.31167	USGS 01394500	AN0194	1-Aug-17	913	7.6	1.3	0.12	Poor
Rahway River at Rahway	R8	40.61889	-74.28333	USGS 01395000	AN0195	1-Aug-17	746	7.9	1.3	0.1	Poor
Ramapo River	Ra	41.09806	-74.16278	USGS 01387500	AN0266	1-Aug-17	845	7.6	1.7	0.1	Poor
Saddle River	S	40.89028	-74.08056	USGS 01391500	AN0290	1-Aug-17	1169	8.2	5.1	0.9	Poor

Table 2: PERMANOVA assemblage ~ site model performance. EukOTU = Eukaryote OTU data set (n = 5,869); DtOTU = Diatom-only OTU data set (n = 359); ChlorOTU = Chlorophyta-only OTU data set (n = 366); HolOTU = Holozoa-only OTU data set (n = 750); SAROTU = SAR_other-only OTU data set (n = 2,070); DtCount = diatom morphological count data set (n= 236).

Model	Df	SumsOfSqs	MeanSq	F	R^2	p-value
EukOTU ~ site	13	17.0439	1.31107	10.515	0.66133	<0.001
DtOTU ~ site	13	17.1364	1.31818	25.162	0.82372	<0.001
ChlorOTU ~ site	13	17.6796	1.35997	16.355	0.75232	<0.001
HolOTU ~ site	13	15.204	1.16955	6.726	0.55538	<0.001
SAROTU ~ site	13	16.099	1.23841	7.3569	0.57739	<0.001
DtCount ~ site	13	5.9351	0.05384	8.479	0.79744	<0.001

Table 3. CCA assemblage ~ impairment model performance of metabarcoding and morphology-based data sets. Each CCA is constrained only by site impairment classification. EukOTU = Eukaryote OTU data set (n=5,869); DtOTU = Diatom-only OTU data set (n=359); ChlorOTU = Chlorophyta-only OTU data set (n = 366); HolOTU = Holozoa-only OTU data set (n = 750); SAROTU = SAR_other-only OTU data set (n = 2,070); DtCount = diatom morphological count data set (n = 236). PVE1 – percent variance in species data explained by the first (constrained) CCA axis.

Model	PVE1 (%)	F-ratio	p-value
EukOTU ~ Impairment	5.3	4.6	0.01
DtOTU ~ Impairment	10.3	9.4	0.01
ChlorOTU ~ Impairment	7.6	6.7	0.01
HolOTU ~ Impairment	4.4	3.7	0.01
SAROTU ~ Impairment	4.3	3.7	0.01
DtCount ~ Impairment	11.9	5.4	0.01

H. Figures

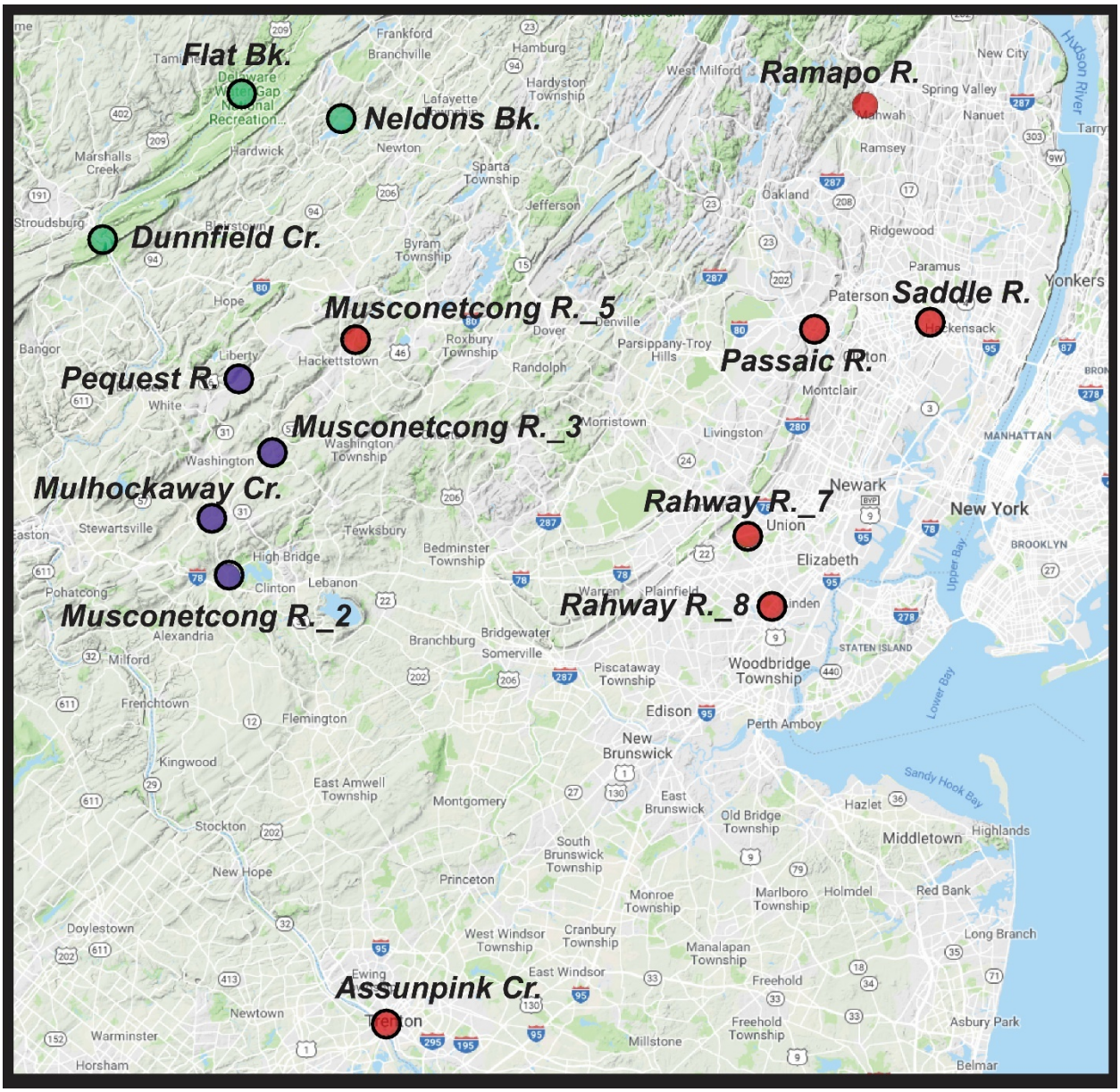

Figure 1. Map of sampling sites. Green - least impaired, purple – moderately impaired, red-severely impaired.

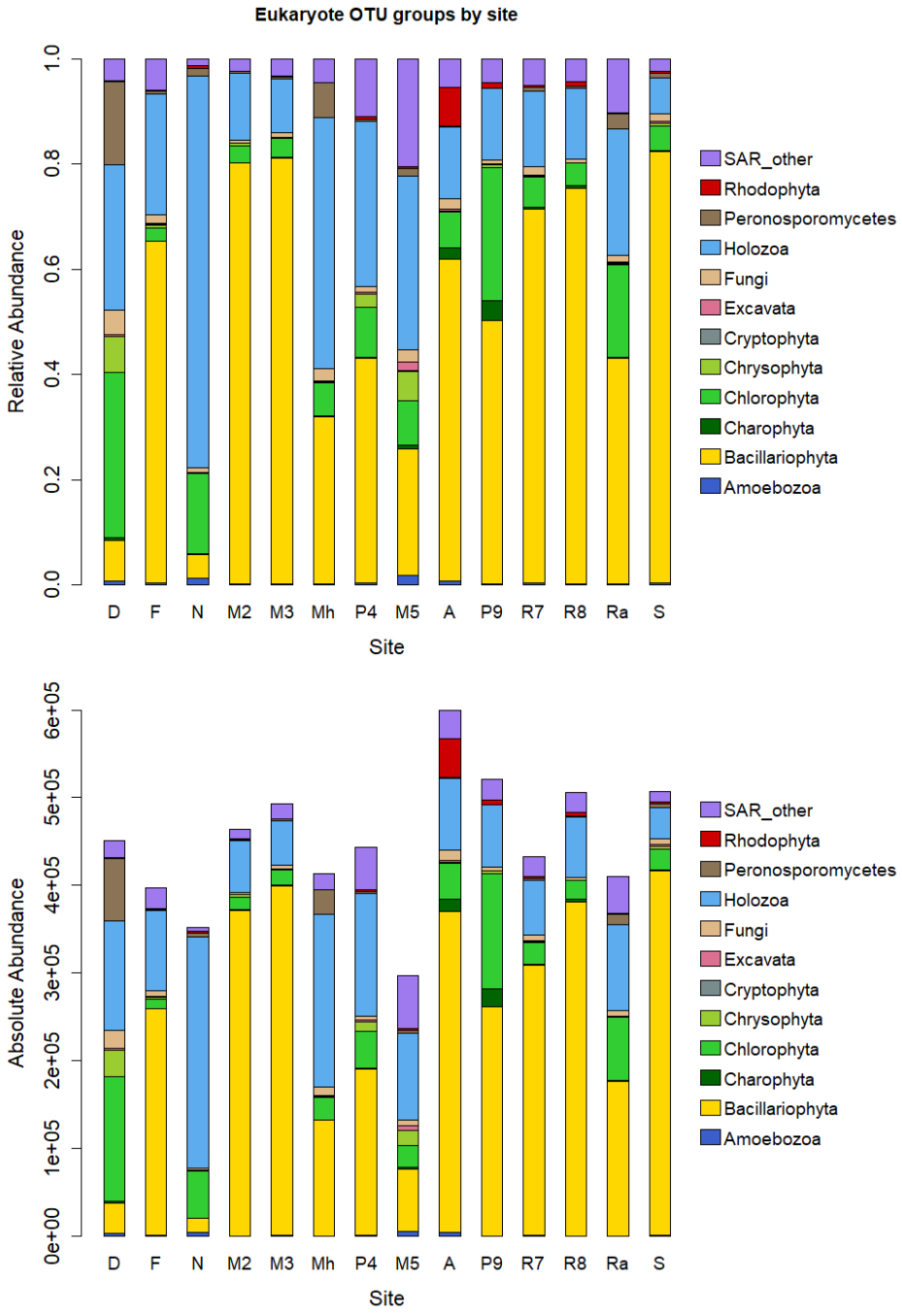

Figure 2. Taxonomic composition of eukaryotic assemblages in studied stream sites, as the percentage (upper plot) and absolute numbers (lower plot) of sequences from major eukaryotic groups. Unassigned sequences are omitted. The dominant groups across the dataset are diatoms (Bacillariophyta), Holozoa, and green algae (Chlorophyta).

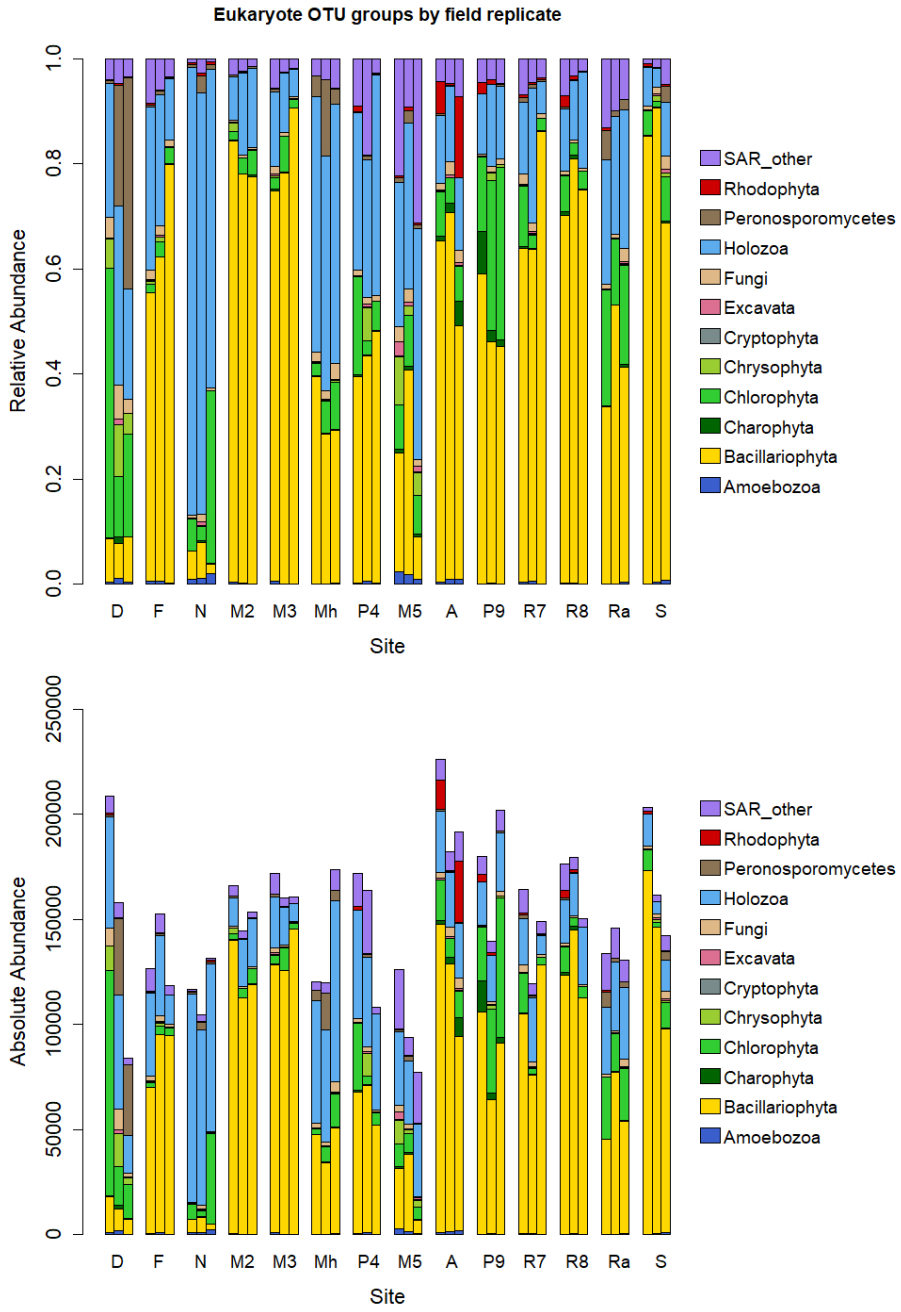

Figure 3. Taxonomic composition of eukaryotic assemblages in field replicate samples, expressed as percentages (upper plot) and absolute numbers (lower plot) of sequences from major eukaryotic groups. Unassigned sequences are omitted. Sequences from two lab replicates for each field replicate were pooled together. Within-site variability is lower than among-site variability.

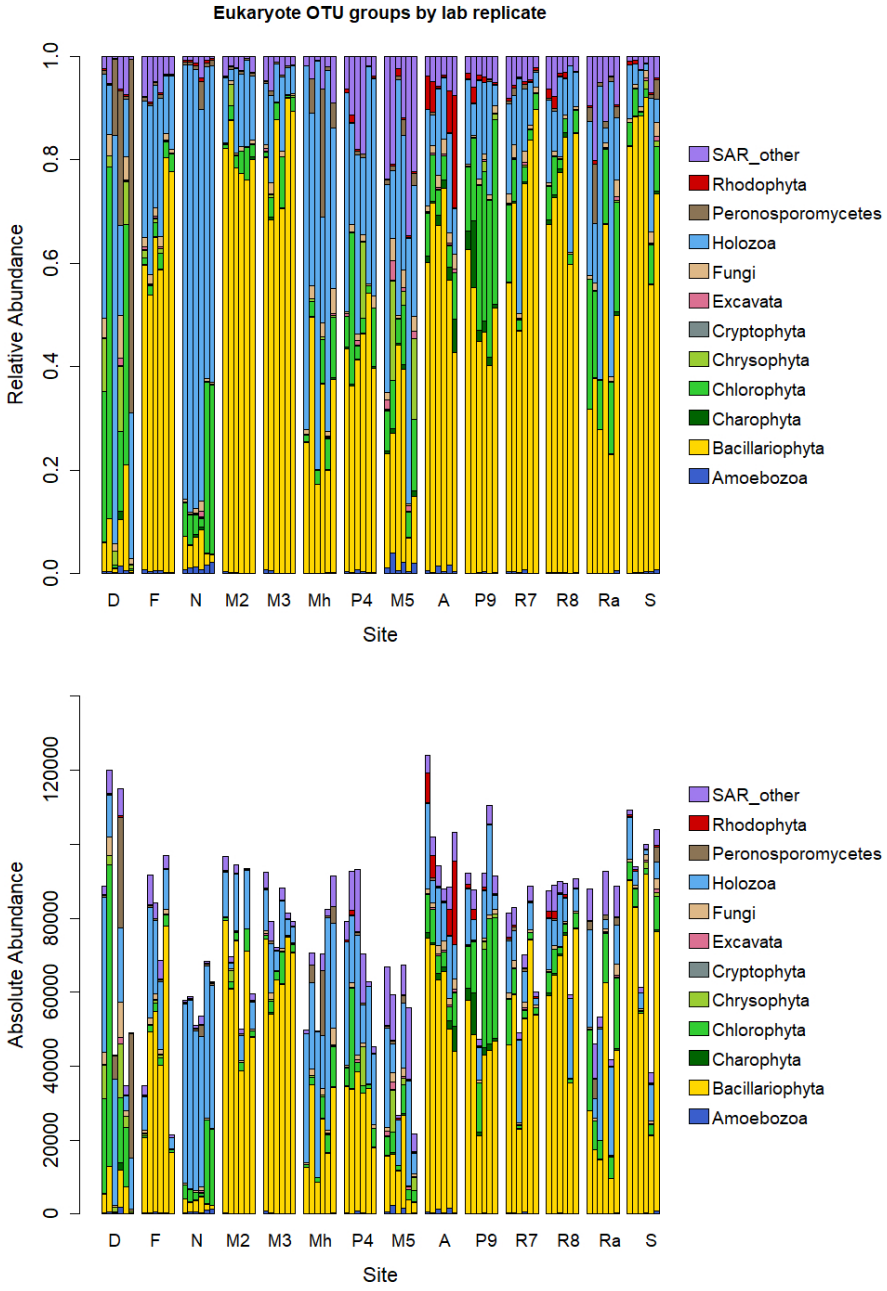

Figure 4. Taxonomic composition of eukaryotic assemblages in 84 stream biofilm samples from 14 stream sites, 3 field replicates per site, 2 lab replicates per field replicate, expressed as the percentage (upper plot) and absolute numbers (lower plot) of sequences assigned to major eukaryotic groups. Unassigned sequences are omitted. Six columns within each site are sorted by field replicate and then by lab replicate, so that the first two columns represent two lab replicates of the first field replicate, the next two columns represent two lab replicates of the second field replicate, etc.

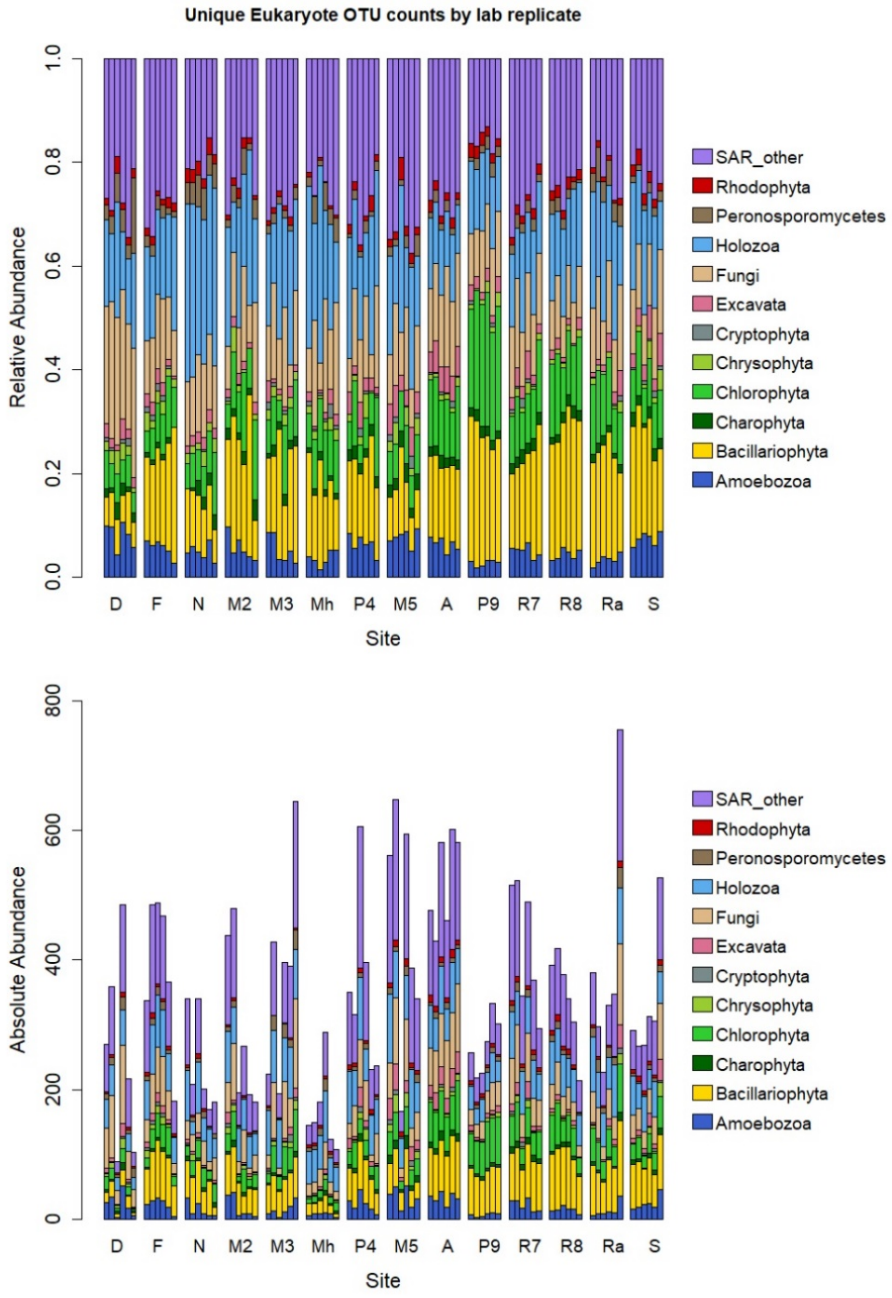

Figure 5. Taxonomic composition of eukaryotic assemblages in 84 stream biofilm samples from 14 stream sites, 3 field replicates per site, 2 lab replicates per field replicate. Abundance is expressed as presence of individual OTUs assigned to major eukaryotic groups, as percentages (upper plot) and absolute numbers (lower plot). Six columns within each site are sorted by field replicate and then by lab replicate, so that the first two columns represent two lab replicates of the first field replicate, the next two columns represent two lab replicates of the second field replicate, etc.

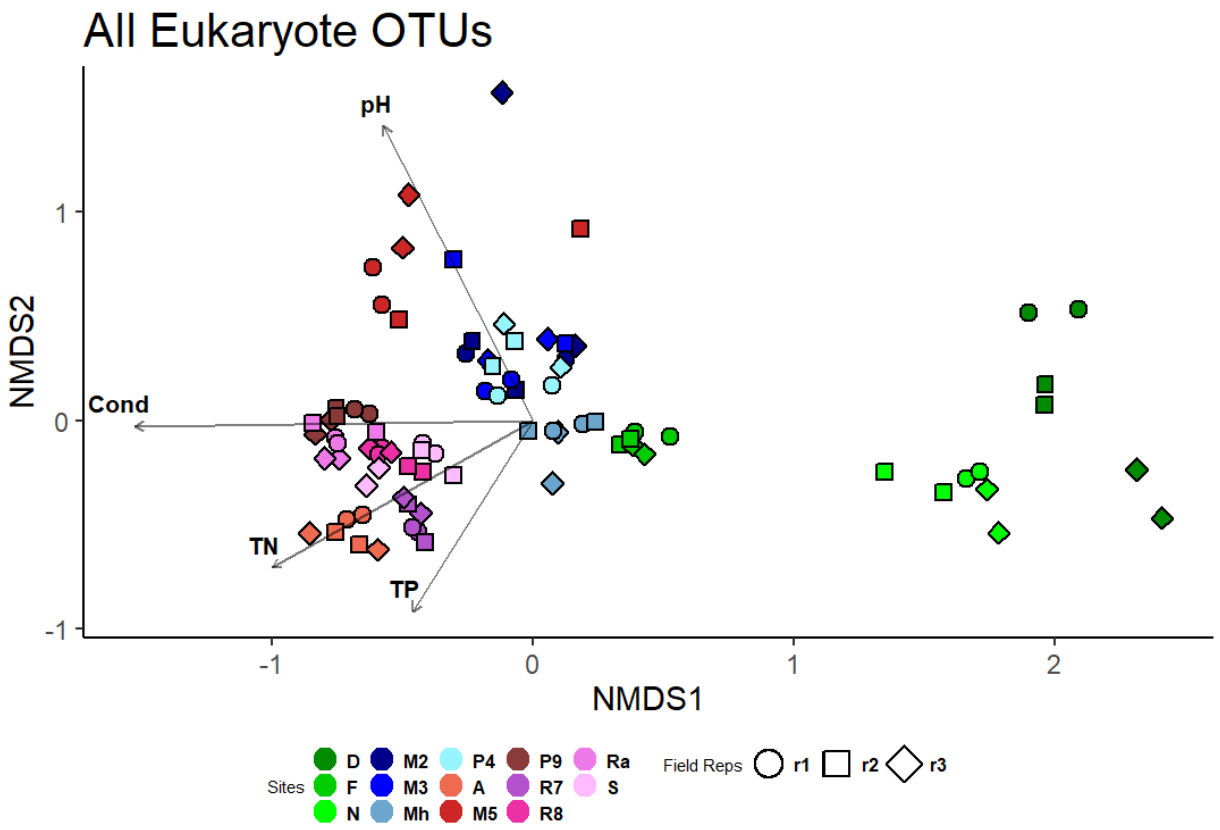

Figure 6. NMDS plot of all 5,869 eukaryotic OTUs and fitted environmental variables. Relative abundance data were Hellinger-transformed. Shades of green– samples from the least impaired sites; shades of blue– samples from the moderately impaired sites; shades of red– samples from the heavily impaired sites. Cond. – electrical conductivity; TP – Total Phosphorus, TN – Total Nitrogen.

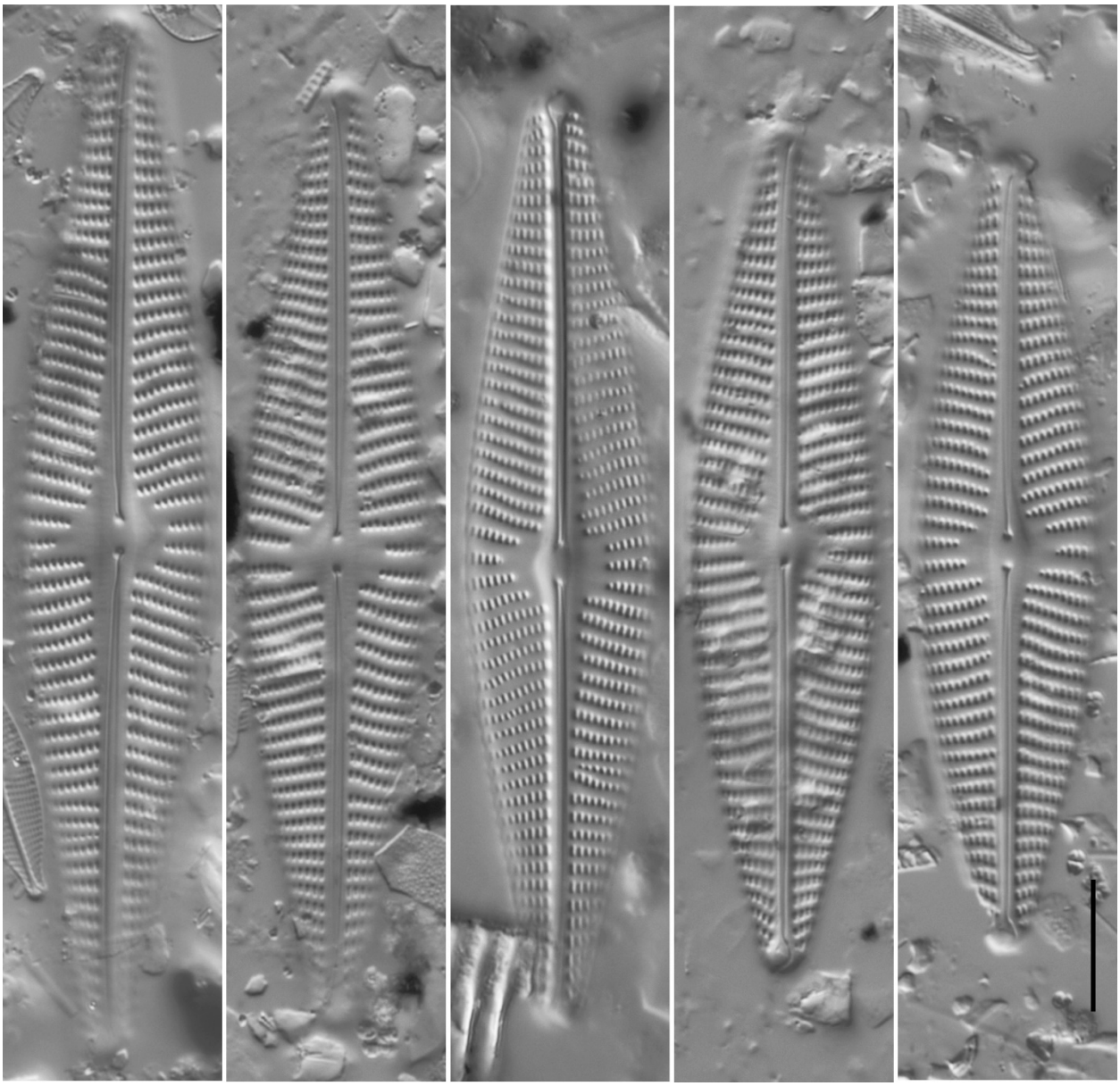

Figure 7. Selected specimens of *Navicula peregrinopsis* from studied streams. This species is reported for the first time from the State of New Jersey; its populations appear to expand in response to increased salinization of rivers. Scale bar = 10 μm.

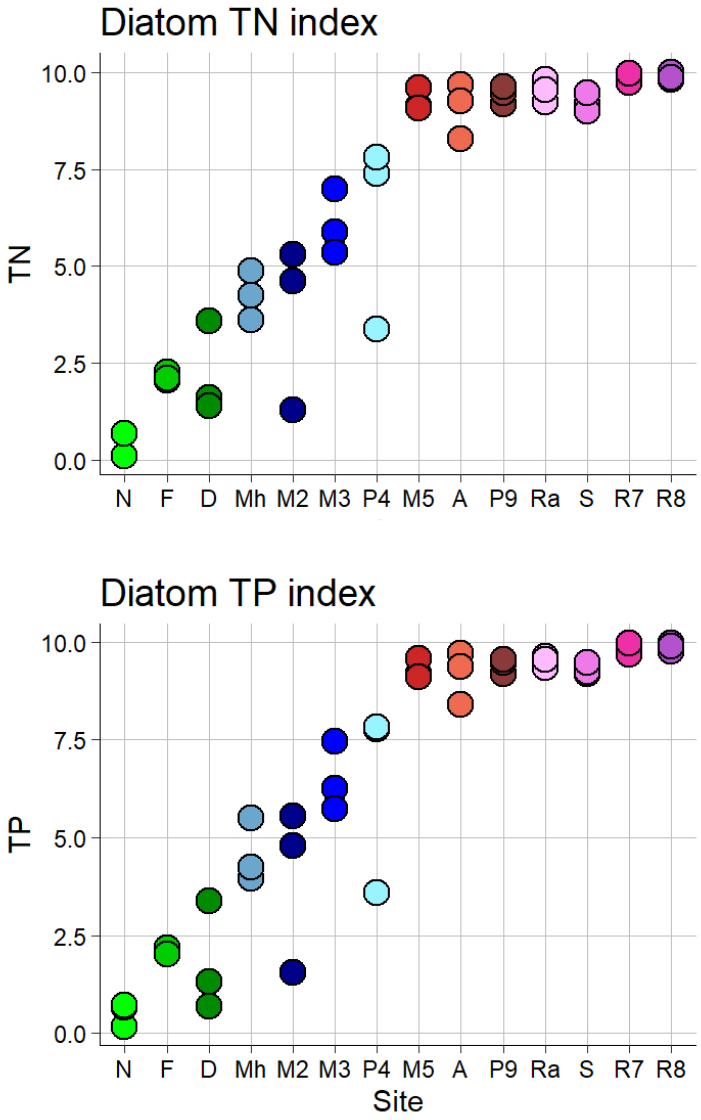

Figure 8. Diatom nutrient indices (Potapova & Charles 2007) calculated for 42 samples from 14 New Jersey rivers. Green circles represent minimally impaired sites, blue circles – moderately impaired sites and red circles – heavily impaired sites. Index values vary from 0 (no nutrient impairment) to 10 (extremely high nutrient impairment).

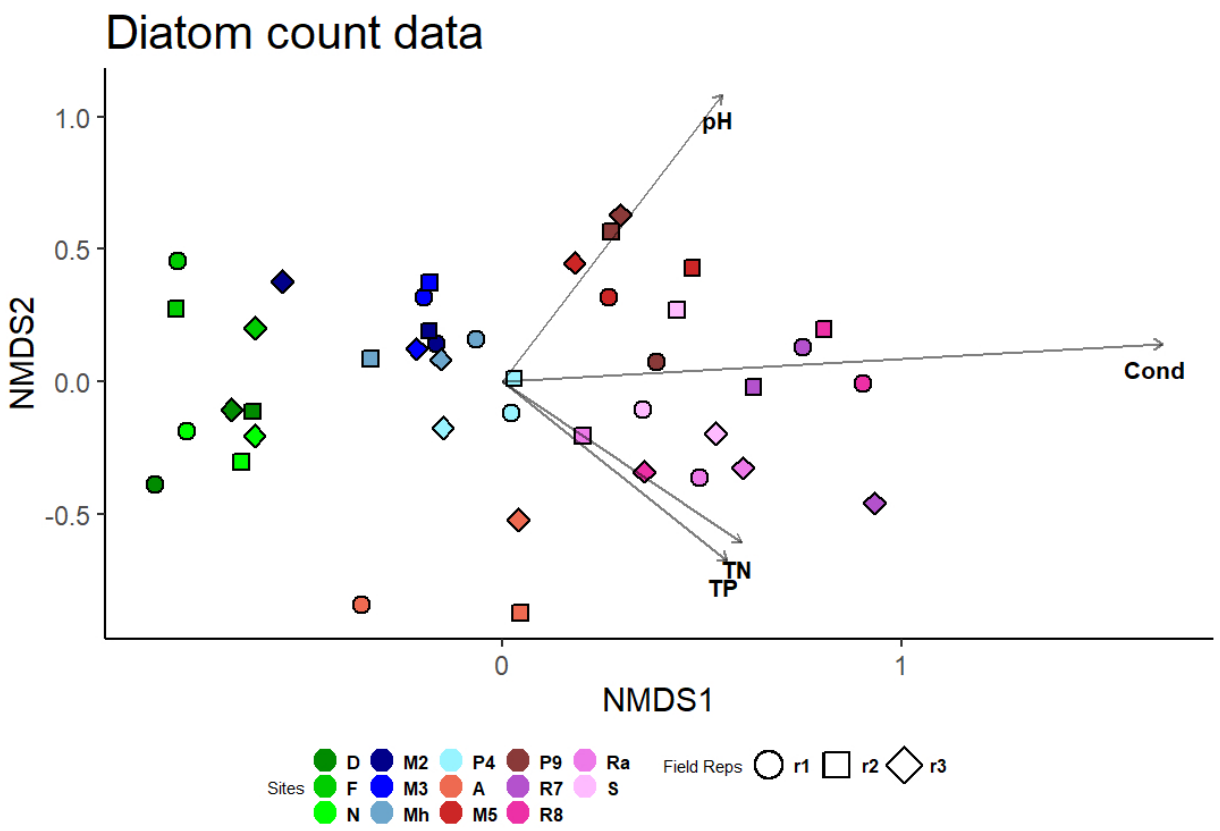

Figure 9. NMDS ordination of diatom count data: plot showing position of samples and fitted vectors of environmental variables. Green symbols – samples from the least impaired sites; blue symbols – samples from the moderately impaired sites; red symbols – samples from the heavily impaired sites. Various shades represent different sites, while three different symbols (circles, squares and diamonds) represent three field replicates. Cond. – electrical conductivity; TP – Total Phosphorus, TN – Total Nitrogen.

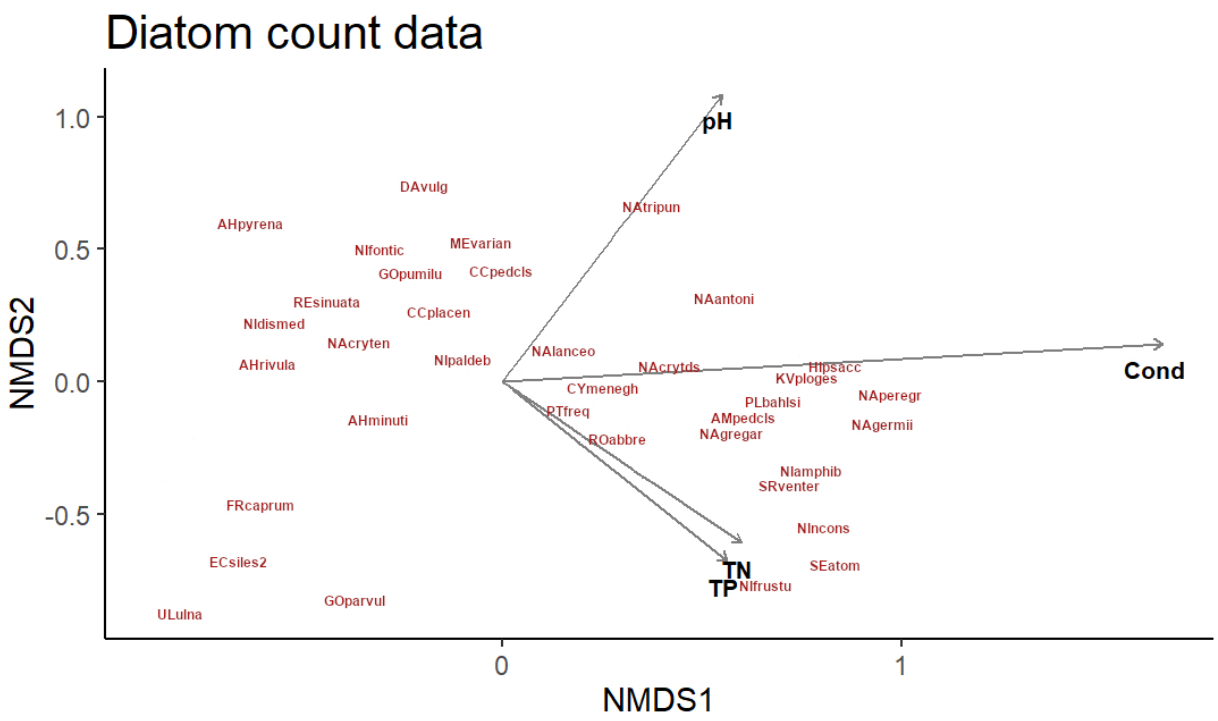

Figure 10. NMDS ordination of diatom count data: plot showing position taxa and fitted vectors of environmental variables. Only taxa that reached 10% relative abundance in at least one sample are identified by “short codes” in red text. Taxon “short names” are in Appendix 2.

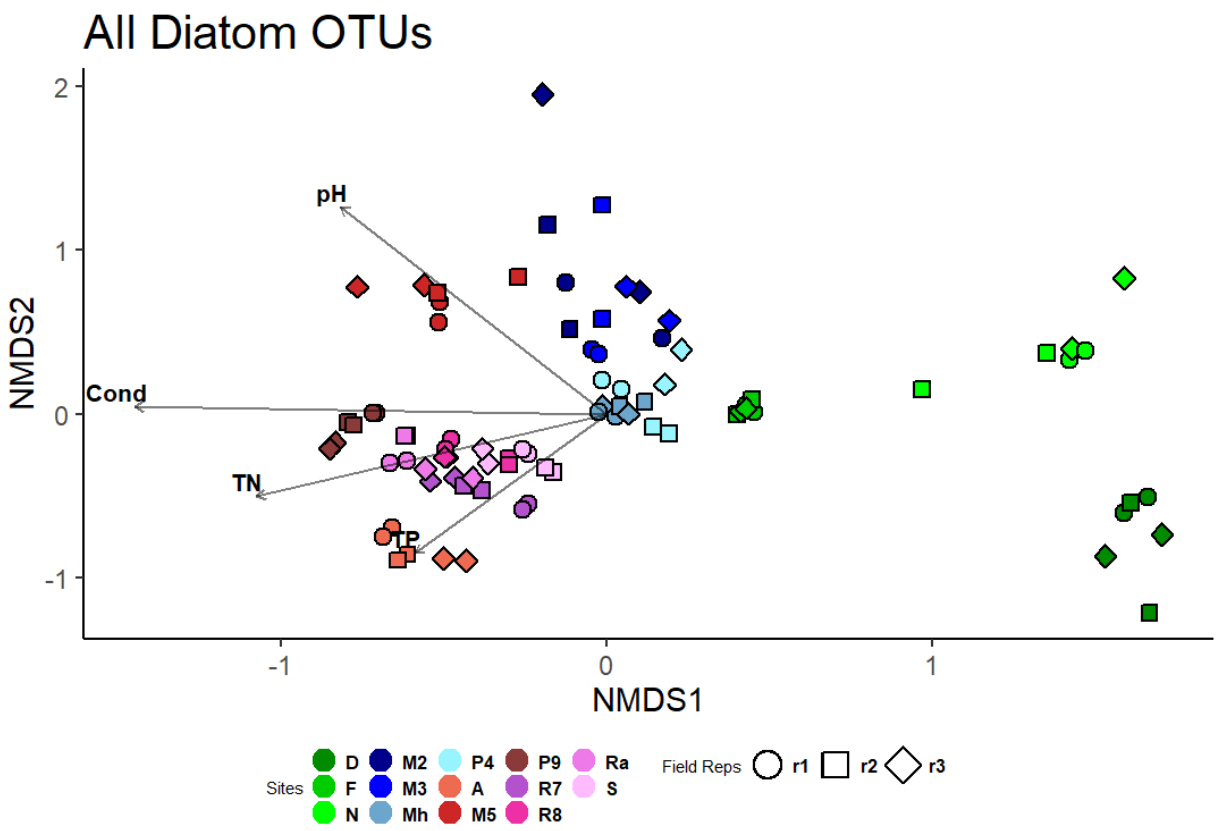

Figure 11. NMDS ordination of 359 diatom molecular OTUs: plot showing position of samples and fitted vectors of environmental variables. Green symbols – samples from the least impaired sites; blue symbols – samples from the moderately impaired sites; red symbols – samples from the heavily impaired sites. Various shades represent different sites, while three different symbols (circles, squares and diamonds) represent three field replicates. Cond. – electrical conductivity; TP – Total Phosphorus, TN – Total Nitrogen.

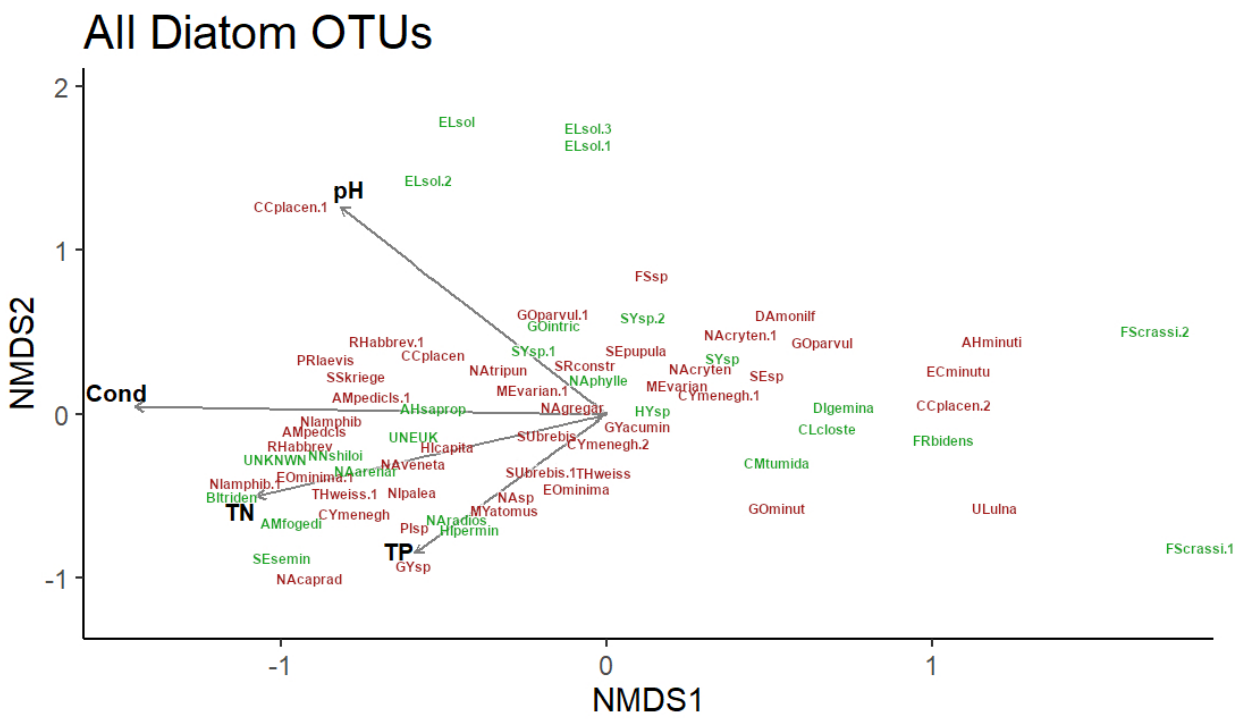

Figure 12. NMDS ordination of 359 diatom molecular OTUs: plot showing position of taxa and fitted vectors of environmental variables. Only OTUs that reached 20% relative abundance in at least one sample are shown. The OTUs assigned to taxa also found in microscope counts are labeled by red text, while those that did not appear in counts are labeled by green text. OTU labels are “short names” available in Appendix 1.

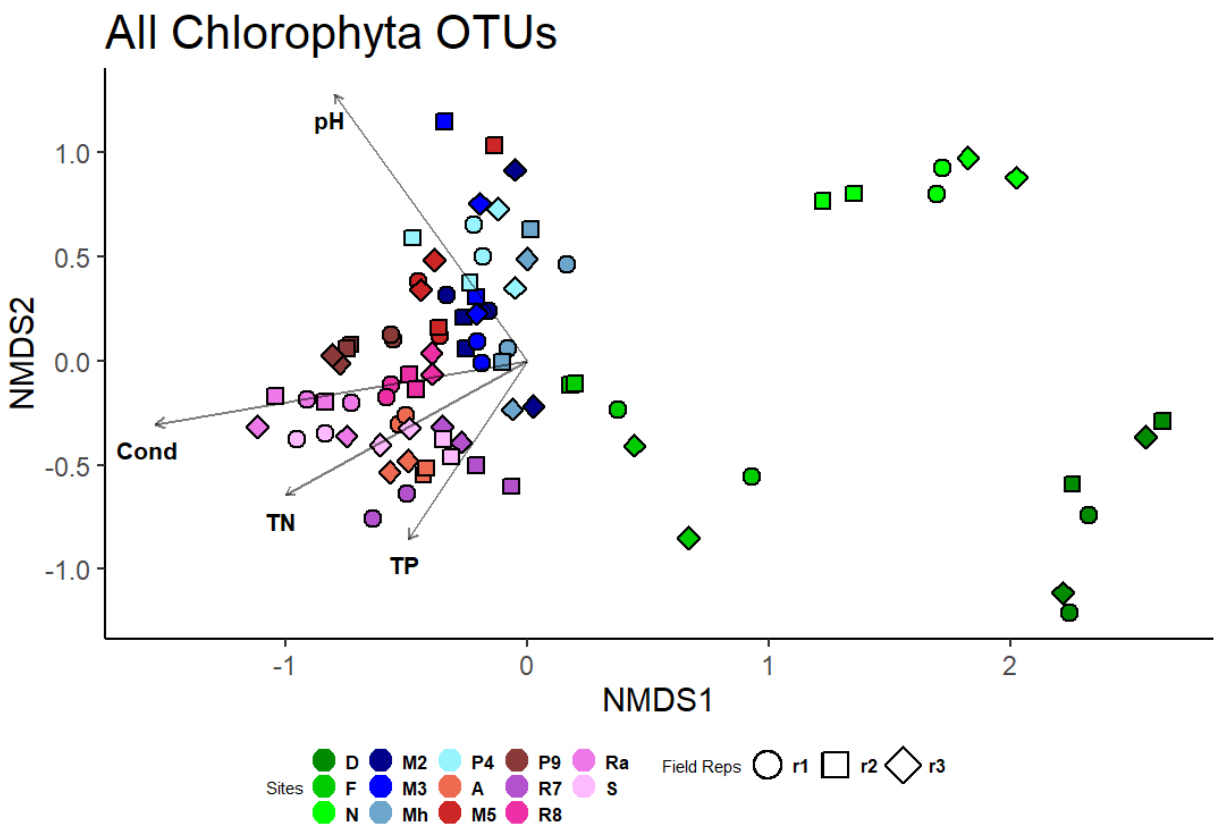

Figure 13. NMDS ordination of 366 Chlorophyta molecular OTUs: plot showing position of samples and fitted vectors of environmental variables. Green symbols – samples from the least impaired sites; blue symbols – samples from the moderately impaired sites; red symbols – samples from the heavily impaired sites. Various shades represent different sites, while three different symbols (circles, squares and diamonds) represent three field replicates. Cond. – electrical conductivity; TP – Total Phosphorus, TN – Total Nitrogen.

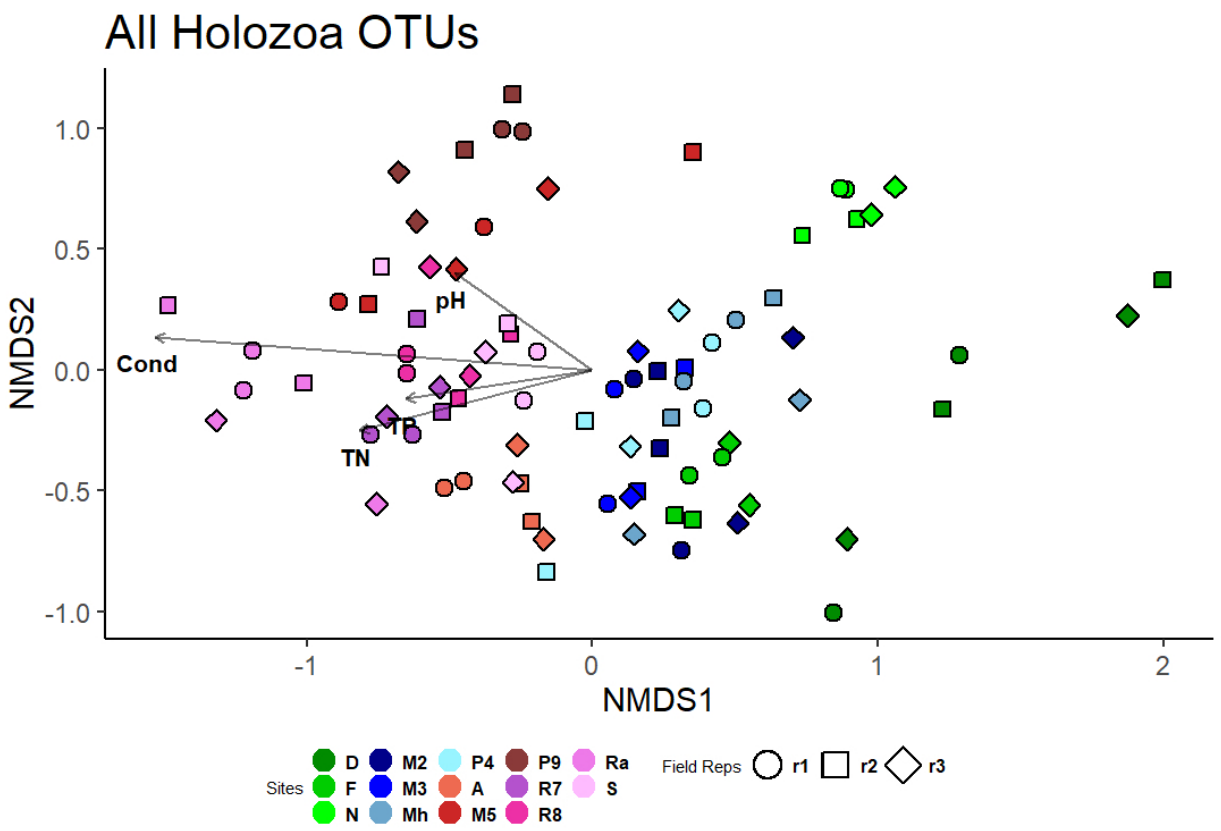

Figure 14. NMDS ordination of 750 Holozoa molecular OTUs: plot showing position of samples and fitted vectors of environmental variables. Green symbols – samples from the least impaired sites; blue symbols – samples from the moderately impaired sites; red symbols – samples from the heavily impaired sites. Various shades represent different sites, while three different symbols (circles, squares and diamonds) represent three field replicates. Cond. – electrical conductivity; TP – Total Phosphorus, TN – Total Nitrogen.

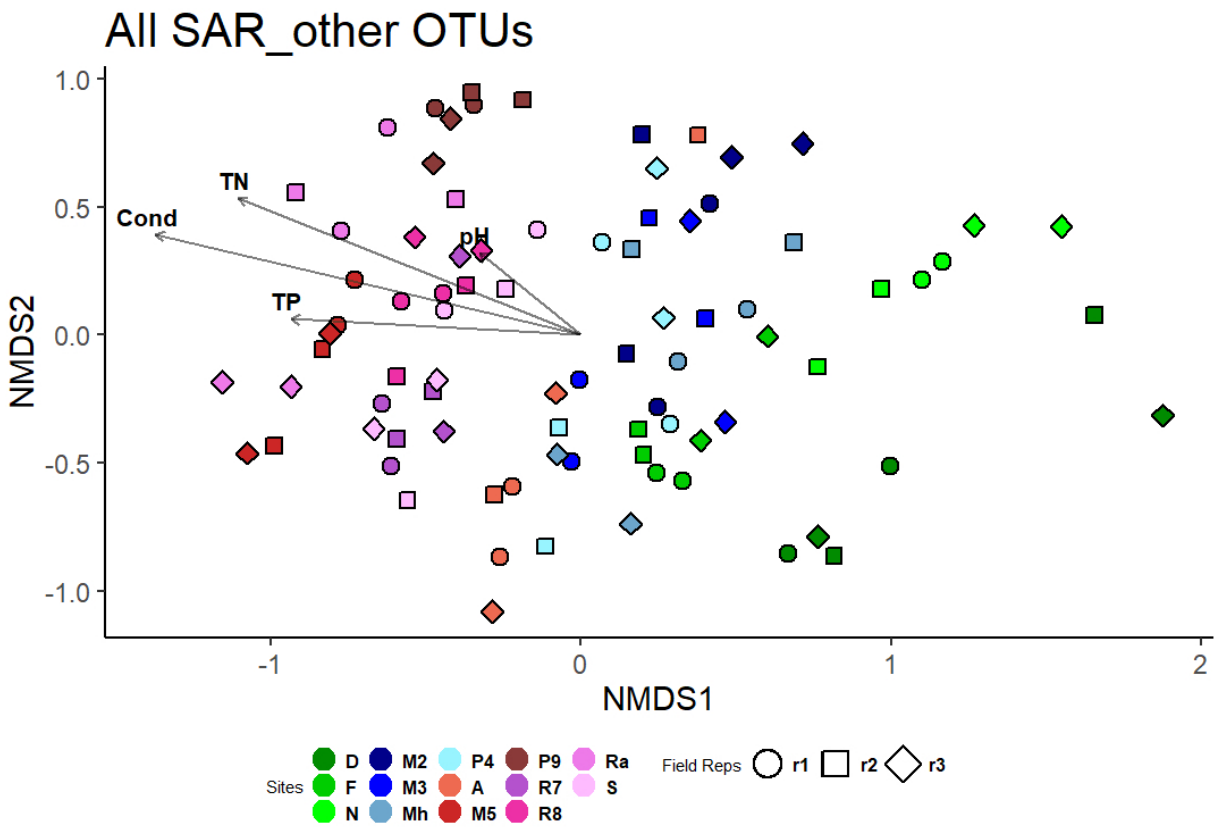

Figure 15. NMDS ordination of 2,070 "SAR_other" molecular OTUs: plot showing position of samples and fitted vectors of environmental variables. Green symbols – samples from the least impaired sites; blue symbols – samples from the moderately impaired sites; red symbols – samples from the heavily impaired sites. Various shades represent different sites, while three different symbols (circles, squares and diamonds) represent three field replicates. Cond. – electrical conductivity; TP – Total Phosphorus, TN – Total Nitrogen.

I. Appendices

APPENDIX 1. Excel file with metabarcoding data: a matrix of unique sequences with their abundance in each sample.

APPENDIX 2. Excel file with diatom species data, counts.