Identification of diatom DNA barcodes for biomonitoring of the New Jersey Pine Barrens aquatic ecosystems

Authors and Affiliation:

Dr. Marina Potapova
Academy of Natural Sciences of Drexel University

Prepared for:
New Jersey Department of Environmental Protection
Division of Science and Research
Project Manager: Mihaela Enache

August 18, 2024

Funded by the New Jersey Department of Environmental Protection
through an agreement with the Sea Grant Consortium
Contract Number: SR21-013

Acknowledgements:

The New Jersey Pinelands Commission (NJPC) provided support with pond selection and environmental data. Special thanks to NJPC's Patrick Burritt for helping with project fieldwork including diatometer setup and collection.

Please cite as:

Potapova, M. 2024. Identification of diatom DNA barcodes for biomonitoring of the New Jersey Pine Barrens aquatic ecosystems. Final Report to the NJ Department of Environmental Protection. Trenton, NJ. 40 Pages. https://hdl.handle.net/10929/140480

TABLE OF CONTENTS

Executive summary

The goal of this project was to investigate the use of DNA metabarcoding for evaluating and monitoring environmental health of waterbodies in the New Jersey Pinelands. Metabarcoding is a technique of taxonomic identification of organisms in environmental samples via analysis of short DNA sequences. We explored metabarcoding of diatom and other protistan assemblages in ephemeral ponds which are increasingly appreciated as valuable ecosystems and targets for environmental conservation. Metabarcoding has proven to be an efficient approach to monitor changes in aquatic biological communities circumventing the need for time and labor-consuming visual identification of organisms. While this approach has many advantages, such as reduced cost and the ability to characterize multiple taxonomic groups simultaneously, it also has some shortcomings. The major limitation of metabarcoding is incompleteness of the taxonomic reference databases that leaves many DNA sequences unassigned to taxa. For example, the reference database suitable for diatom metabarcoding (Diat.barcode) was developed mostly using diatoms from European rivers and therefore, does not have a good coverage of other geographic areas and habitats. Adding new records to this database typically involves culturing and sequencing diatoms, which is often prohibitively expensive. In this project we explored an opportunity to establish diatom barcodes from natural samples with low species diversity to establish correlation between dominant morphospecies and most abundant sequences.

In 2021-2023 the New Jersey Pinelands Commission and the New Jersey Department of Environmental Protection (NJDEP) staff installed diatometers which are floating devices holding microscope slides that serve as artificial substrates for diatom colonization, in several ponds in the New Jersey Pine Barens. Some natural substrates were sampled concurrently with diatometers in the same ponds in 2022-2023. We used samples from these natural and artificial substrates for metabarcoding employing two sets of primers, one covering a wide variety of microbial eukaryotes (18S_V9) and another targeting diatoms (a fragment of *rbc*L gene). Our aims were to assess the effectiveness of metabarcoding for characterization of microbial eukaryotic and diatom assemblages and to explore the opportunity for establishing novel barcodes for diatoms that are not represented in reference databases.

Both sets of primers retrieved hundreds of unique DNA sequences (amplicon sequence variants or ASVs usually corresponding to biological species) and millions of reads (individual DNA fragments representing individual organisms), covering all main target groups of organisms and revealing a vast diversity of eukaryotic microorganisms. We found that assemblages assessed both microscopically and by metabarcoding responded to water-quality differences among the ponds. We were also able to establish correspondence between eleven diatom morphotaxa and *rbc*L amplicon sequences obtained via metabarcoding, thus providing data useful for future monitoring of aquatic ecosystems in New Jersey. While only a small portion of morphotaxa could be related to DNA barcodes, some of these taxa are extremely abundant in studied ponds and the new assignments raise the portion of sequences that could be assigned taxonomically from 21 to 50%. We conclude that metabarcoding is an extremely valuable tool for biological monitoring of New Jersey waterbodies even in the absence of complete reference databases as sequences themselves can serve as environmental indicators. Further investigations are likely to improve the quality of taxonomic assignments either by establishing barcodes from environmental samples as was done in this project, or by standard culturing and Sanger sequencing approaches.

1. Introduction

The goal of this project was to explore the use of diatom DNA metabarcoding for evaluating environmental health of waterbodies in the New Jersey Pinelands. While river and lake biota understandably has been a focus of biodiversity and bioassessment studies, ephemeral water bodies are recently drawing considerable attention as essential landscape features valuable for environmental conservation (Barta et al. 2023, Hill et al. 2021). Coastal-plain ponds are important landscape features of the New Jersey Pine Barrens providing habitat for native plant and animal species and thus contributing to regional biodiversity (Bunnell et al. 2018). New Jersey DEP previously conducted a survey of pond biota focused on larger charismatic organisms, while this is the first project dealing with microbial and algal assemblages in Pinelands ponds.

The diatom flora of ephemeral ponds of New Jersey Pinelands has not been previously studied, but several diatom surveys have been conducted in the past on diatoms from rivers and lakes of this region. Zampella et al. (2007) investigated diatoms in blackwater streams, reporting 132 taxa from diatometers installed at 14 sampling sites and citing insufficient knowledge of diatom diversity and ecology as the main impediment to successful use of diatom as bioindicators in this setting. Ponader et al. (2008) likewise, focused their study of diatoms from running waters, but covering larger area in New Jersey coastal plain, including a few sites in Pinelands. Their goal was to construct diatom inference models for monitoring nutrient pollution, and therefore the diatom assemblage composition was only documented by a list of common species. The most comprehensive floristic data on diatoms from New Jersey Pinelands are found in Siver & Hamilton (2011) who documented diatom occurrence in lakes across the Eastern US seaboard including several lakes in Pinelands by extensive collection of light and electron microscopy imaging. This project aims at characterizing diatom assemblages in ponds using both traditional microscopy and DNA metabarcoding that emerges as a promising bioassessment tool.

Metabarcoding has proven to be an efficient approach to monitor changes in aquatic biological communities circumventing the need for time and labor-consuming visual identification of organisms (Pawlowski et al. 2018). While this approach has many advantages, such as reduced cost and the ability to characterize multiple taxonomic groups simultaneously, it also has some shortcomings (Keck et al. 2017). The major limitation of metabarcoding is incompleteness of the taxonomic reference databases that leaves many DNA sequences unassigned to taxa (Rimet et al. 2018). For example, the reference database suitable for diatom metabarcoding (Diat.barcode) was developed mostly using diatoms from European rivers and therefore, lack information on diatoms

from other geographic areas and habitats. Adding new records to this database involves culturing and sequencing diatoms, which is often prohibitively expensive. In this project we explored an opportunity to establish diatom barcodes from natural metabarcoded samples where species diversity is low and therefore, there is an opportunity to establish correlation between dominant morphospecies and most abundant sequences. This approach has been successfully used by Rimet et al. (2018) and Kochoska et al. (2023) who added several sequences to reference databases by analyzing correspondence between sequencies obtained by metabarcoding and diatoms identified microscopically and even described new species using these data.

In 2021-2023 the New Jersey Pinelands Commission and NJDEP staff installed diatometers which are floating devices holding microscope slides that serve as artificial substrates for diatom colonization, in several ponds. In addition, some natural substrates were sampled concurrently with diatometers in the same ponds in 2022-2023. We used biofilm samples from these natural and artificial substrates for DNA metabarcoding employing two sets of primers, one covering a wide variety of microbial eukaryotes (18S_V9) and another targeting diatoms (a fragment of *rbc*L gene). Our aims were twofold: first to assess the effectiveness of metabarcoding for characterization of microbial eukaryotic and diatom assemblages in Pinelands ponds, and second to explore the opportunity for establishing novel barcodes for diatoms that are not represented in reference databases.

2. Methods

2.1. Sampling

A total of 61 samples from 30 ponds were used in this project (Table 1, Fig. 1). In 2021, 17 diatometer samples were collected from diatometers installed in 17 ponds. The 2022-2023 sites were selected based on diatom count results from the 40 samples collected in 2018-2019 and analyzed by M. Enache. In 2022, 21 samples (including 6 diatometer slides) were collected from 6 ponds, and in 2023, 23 samples (including 7 diatometer slides) were collected from 8 ponds. Most sampled ponds were very small and known to dry by the mid- or late summer. Some ponds were natural and some excavated. All samples were frozen and initially stored at -20^0 C, with later transfer to the -80^0 C storage at the Academy of Natural Sciences of Philadelphia (ANS). Water-quality data were collected in 2017 (Bunnell et al. 2018) and provided by the New Jersey Pinelands Commission.

Table 1. The list of sampled ponds, samples, and water quality characteristics used in the analysis. The absence of data is denoted by an "nd". The number of diatometers, sediment samples, and plant surveys are denoted by 0, 1, or 2.

Site Name	Longitude DD	Latitude DD	Chlorophyll A (µg/l)	pH	Specific conductance (µS/cm)	NH3 +NOx (mg/L N)	Orthophosphate (mg/L P)	Chloride (mg/L)	Sampling Year	Diatometer	Sediment	Plants
Basket Pond	-74.866784	39.272322	39.4	4.10	51.5	nd	0.004	6.5	2021	1	0	0
Cardinal Basin	-74.91384	39.70239	13.2	5.95	42.0	0.070	0.004	3.4	2021	1	0	0
Cedar Basin	-74.881036	39.535637	9.8	7.00	34.0	0.010	nd	3.1	2021	1	0	0
Corkery Basin	-74.98614	39.66943	17.8	6.05	59.0	0.015	0.034	4.4	2021	1	0	0
Country Basin	-74.92196	39.69966	50.5	5.85	42.0	0.010	0.004	1.5	2021	1	0	0
Deer Pond	-74.904005	39.378016	7.2	3.85	135.5	0.005	nd	28.8	2021	1	0	0
English Pond	-74.624342	39.409455	42.4	4.85	29.0	0.355	nd	3.1	2021	1	0	0
Fletcher Basin	-74.92261	39.70959	225.3	5.55	29.5	0.005	0.021	1.4	2021	1	0	0
Forbidden Pond	-74.778476	39.707119	14.9	4.35	24.0	nd	nd	2.7	2021	1	0	0
Leah Basin	-74.62765	39.46049	18.1	6.15	40.5	0.040	nd	4.0	2021	1	0	0
MacDonald XPond	-74.863075	39.431816	897.1	5.65	124.0	nd	0.004	14.4	2021	1	0	0
MacKay XPond	-74.837293	39.257774	22.6	5.00	36.0	nd	0.004	3.3	2021	1	0	0
Muirfield Basin	-74.92621	39.70522	292.3	5.30	49.0	0.010	0.038	3.6	2021	1	0	0
Pennypot Pond	-74.850400	39.618662	146.5	4.00	43.5	nd	nd	2.8	2021	1	0	0
Robbins Pond	-74.875740	39.643569	3.5	4.10	37.0	nd	nd	2.6	2021	1	0	0
Third XPond	-74.802565	39.481877	17.3	4.65	27.0	nd	nd	2.6	2021	1	0	0
Windsor Basin	-74.59257	39.43299	20.0	6.10	57.0	0.015	0.017	3.4	2021	1	0	0
Cooper Pond	-74.908315	39.750191	17.1	4.75	25.5	nd	nd	2.6	2022	1	1	1
Evans XPond	-74.537947	39.678000	57.0	4.80	191.5	0.015	nd	51.3	2022	1	1	1
Goober XPond	-74.685265	39.522344	7.2	5.70	36.5	nd	nd	4.1	2022	1	1	1
Island XPond	-74.415986	39.879643	13.1	4.85	20.5	0.010	nd	2.9	2022	1	1	2
Mosquitofish XPond	-74.897635	39.690702	21.1	5.25	116.5	0.005	nd	26.4	2022	1	0	2
Slab Basin	-74.857630	39.833533	57.5	6.60	76.0	nd	0.016	60.2	2022	1	2	2
Arrowhead Pond	-74.835080	39.733398	11.1	4.10	32.5	0.025	0.030	2.3	2024	0	1	1
Corn Pond	-74.442848	39.646583	7.5	4.00	57.5	nd	0.004	6.8	2024	1	1	1
Flittertown Pond	-74.834092	39.666844	25.1	4.00	38.5	nd	nd	2.7	2024	1	1	1
Goober XPond	-74.685265	39.522344	7.2	5.70	36.5	nd	nd	4.1	2024	1	1	1
Hampton XPond	-74.687845	39.765791	10.6	4.45	14.5	nd	nd	1.9	2024	1	1	1
Holly Pond	-74.664959	39.495383	30.1	4.10	360.0	nd	nd	93.6	2024	1	1	1
Price Pond	-74.822091	39.714239	7.8	3.95	42.5	0.005	0.006	4.1	2024	1	1	1
Wesickaman Pond	-74.770411	39.781214	146.8	4.55	37.0	0.020	0.194	2.5	2024	1	1	1

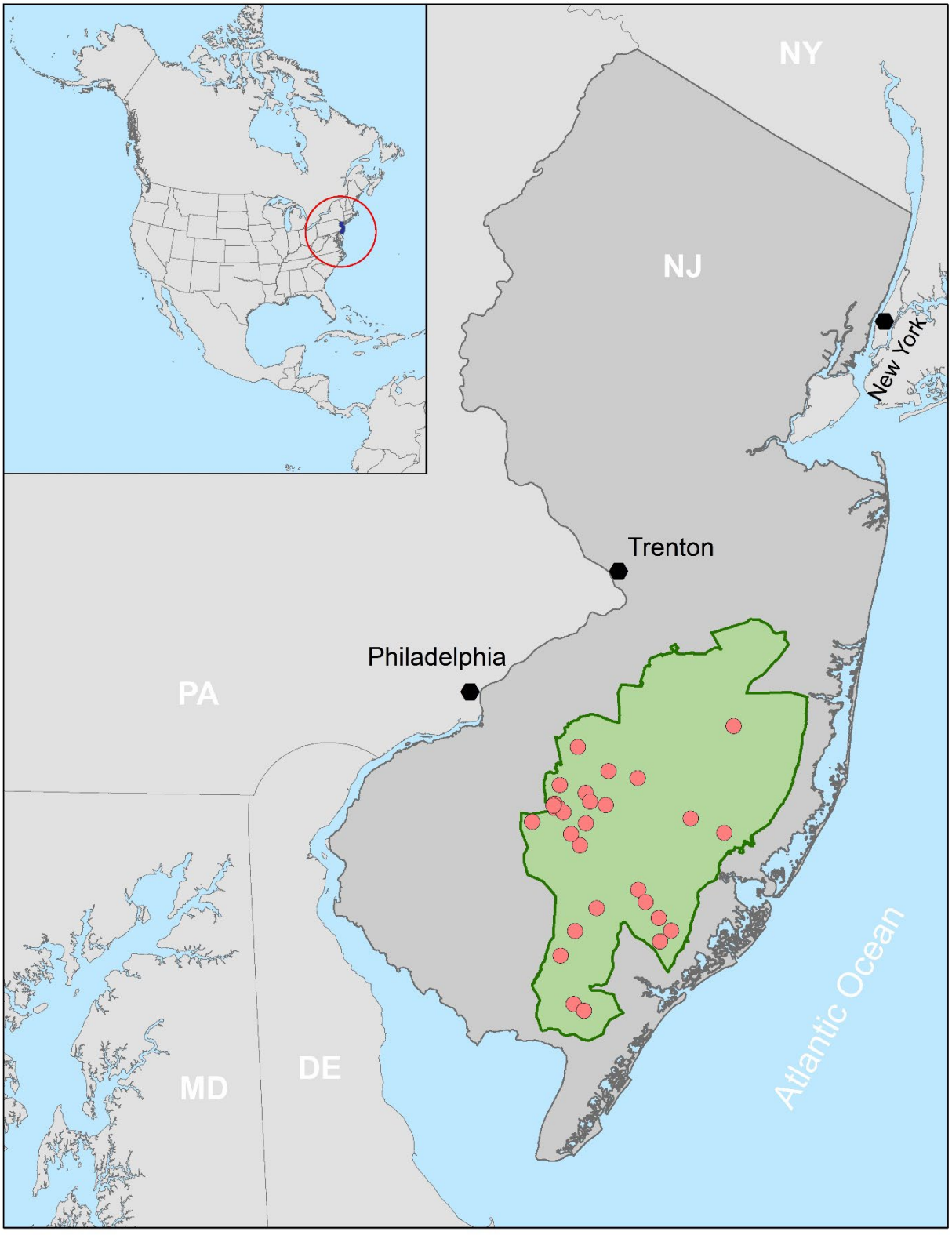

Figure 1. Map showing locations of sampled ponds within the New Jersey Pinelands Area.

2.2. Laboratory processing

The biofilms were scraped from diatometer slides with sterilized razor blades and the slurries were pelleted by centrifugation. These pellets and portions of sediments and aquatic plant samples were then used both for DNA extraction and for preparing permanent diatom slides.

Diatom subsamples were treated with nitric acid, followed by six rinses with distilled water to remove organic matter. Permanent slides were prepared with Naphrax® mounting medium.

2.3 Metabarcoding

Genomic DNA was extracted with Takara NucleoSpin Soil DNA extraction kit (Takara Bio Inc., Shiga, Japan) following the manufacturer's instructions. The DNA was extracted from the total of 61 samples from sediment, submerged aquatic plants, and diatometer slides. DNA yield was quantified with Qubit 3.0 (Invitrogen, Life Technologies, Grand Island, New York, US) The total number of samples used for metabarcoding was 54.

Two markers were used for metabarcoding, the 96–134 base pair (bp) V9 region of the 18S rRNA gene (18S_V9) and a 312 bp fragment of the *rbc*L plastid gene most often used for diatom metabarcoding (Kermarrec et al. 2013, Turk Dermastia et al. 2023, Vasselon et al. 2017, Thompson et al. 2017). The 18S_V9 region was amplified using the 'pan-eukaryotic" primers 1391F, and EukB widely used in protist metabarcoding (Stoeck et al. 2010, Thompson et al. 2017). 18S_V9 libraries were constructed, and sequencing was conducted at the University of Minnesota Genomic Center (UMGC) following protocols by Gohl et al. (2016). The *rbc*L region was amplified with KAPA HiFi HotStart ReadyMix (Kapa Biosystems, Wilmington, MA, USA) and an equimolar mix of the forward primers Diat_ rbcL_708F_1, 708F_2, 708F_3, and the reverse primers R3_1, R3_2 (Vasselon et al. 2017) modified according to the Illumina protocol by adding universal Illumina tails.

The three replicates of each sample were pooled, purified using the AMPure XP Beads (Agencourt Bioscience Corp., Beverly, MA, USA) according to manufacturer's instructions, and sent for the last steps of library construction and sequencing to UMGC. The amplicon libraries were sequenced on an Illumina MiSeq platform using the V2 paired-end sequencing kit (2×300 bp).

2.4. Bioinformatics processing and taxonomic assignment

Illumina paired-end reads were processed using the *dada2* package version 1.18.0 in R version 4.3.1. (R core team 2023). A "read" represents a particular sequence obtained from a DNA fragment from a sample which is used to identify a species present in that sample. Each read would correspond to a single count of that particular species and is obtained directly from the sequencer. Reads can contain sequencing errors and need to be filtered and quality checked. Primer sequences are short nucleic acid sequences (segments of DNA) designed to be complementary to the beginning and the end of the target sequence that will be amplified during the polymerase chain reaction (PCR). Primer sequences were removed with cutadapt version 2.8 (Martin 2011) using the default parameters (maximum error rate = 10%) and the -g flag which removes any base upstream of the primers. Read quality was visualized with the *plotQualityProfile* function. Reads were filtered using the *filterAndTrim* function, adapting parameters (truncLen, minLen, truncQ, maxEE) according to overall sequence quality. Merging of the forward and reverse reads was carried out with the *mergePairs* function using the default parameters (minOverlap = 12, maxMismatch = 0). Chimeras were removed using *removeBimeraDenovo* with default parameters. Parameters in the *dada2* script used for processing *rbcL* reads were those recommended for diatom metabarcoding and available on Github (https://github.com/fkeck/DADA2_diatoms_pipeline).

The SILVA 132 18S reference database (Morien & Parfrey 2018) was used for taxonomic assignment of 18S_V9 unique sequence amplicons (ASV). A unique sequence amplicon (ASV) is a DNA fragment that is amplified and sequenced using a sequencing method called amplicon sequencing. The amplicon sequencing is a technique that uses polymerase chain reaction (PCR) to amplify a specific region of DNA. This method allows for focusing on a specific DNA region of interest instead of sequencing the entire genome. We used the Diat.barcode, version 10 (Rimet et al. 2019) for taxonomic assignment of diatom *rbc*L ASVs. Taxonomic assignment was carried out via *assignTaxonomy* function in *dada2* using 80% as threshold bootstrap values. ASVs matching non-eukaryotic reference sequences and representing predominantly macroscopic groups of organisms were excluded. These included land plants, Rhodophyta (red algae), Pheophyta (brown algae), and most metazoans except Annelida, Gastrotricha, Myxozoa, Nematoda, Plathyhelmintes, and Rotifera.

2.5. Diatom enumeration

Sediment subsamples were treated with nitric acid, followed by six rinses with distilled water to remove organic matter. Permanent diatom slides were prepared with Naphrax® mounting medium and examined under Zeiss AxioImager A1 light microscope equipped with differential contrast optics and oil immersion 100x objective. At least 300 diatom valves were identified and counted in each sample using several identification resources (Spaulding et al. 2021, Krammer & Lange-Bertalot 1986-2004, Lange-Bertalot et al. 2011, Levkov et al. 2016). For 17 samples collected from diatometer slides in 2021, the counts were also provided by Dr. M. Enache; comparison between counts obtained by two analysts were used to assess taxonomic uncertainty of identifications based on microscopy.

2.6. Data analysis

All data manipulations and numerical analyzes were conducted in R environment using packages *vegan* version 2.6-4 (R Core Team 2023) and packages commonly employed for data handling and visualization.

To assess the efficacy of metabarcoding for characterizing biodiversity of microbial eukaryotes and for relating assemblage composition to environmental characteristics, standard multivariate techniques were used. To account for variability in sequencing depth, the raw ASV numbers were rarefied to the smallest number of ASVs per sample in each data subset, sediment or water (*rrarefy* function in *vegan*). Non-metric MultiDimensional Scaling (NMDS) with Hellinger-transformed data and Bray-Curtis distance matrix was used for visualizing patterns in assemblage composition (*metaMDS* function in *vegan*). *Envfit* procedure in *vegan* was employed to explore correlations between NMDS axes and measured environmental characteristics.

To identify *rbc*L barcodes, several species or species complexes that reached relative abundance of at least 50% in at least one sample were selected for examination. These occurrences were compared to the most abundant ASVs in the same samples which were taxonomically assigned to corresponding genera. Evolutionary divergence of sequences was calculated as the number of substitutions per site for sequences assigned to selected most common and diverse genera with the MEGA v.11 program (Tamura et al. 2021). A neighbor-joining tree (Saitou & Nei 1987) was constructed in MEGA using an alignment of 3669 rbcL barcode sequences based on the rbcL sequences provided in the Diat.barcode reference database (version 10.1, Rimet et al. 2021) with added 370 sequences obtained in this project. Portions of this tree that contained

sequences of interest were then examined for the purpose of identifying clusters of sequences potentially belonging to single morphospecies. The second approach for finding correspondence between morphospecies and *rbc*L sequences was to calculate a co-occurrence matrix to find potential associations between morphospecies and ASVs. This was accomplished by computing Bray-Curtis distances using the *vegdist* procedure in *vegan* package in R. Small distances (high co-occurrence) between morphospecies and ASVs do not necessarily indicate their association and may occur because of the "ecological" co-occurrence of species in the dataset; therefore, this approach was only use to screen for potential relationships. The third approach was to identify low-diversity genera in both morphological and rbcL datasets and then to search for correspondence between morphotaxa and ASV occurrences.

3. Results and Discussion

3.1. 18S metabarcoding

18S_V9 sequencing of 61 samples yielded a total of 6,123,792 raw reads. 3,604,113 reads remained after quality filtering, merging, denoising and chimera removal. The total number of ASVs in all 61 samples was 18,086. After removing ASVs that were not taxonomically assigned to Eukaryota with both reference databases, not assigned to any eukaryotic phylum, or representing groups of predominantly multicellular organisms, resulted in a dataset of 1,642,367 reads and 4,022 ASVs (Appendix 1).

Most ASVs belonged to various groups of protists reflecting their rich diversity in the ponds (Fig. 2). Heterotrophic taxa were especially abundant. Ciliates were the most abundant, while cercozoans and ciliates were the most diverse groups. Other abundant heterotrophs were excavates, amoebozoa, peronosporomycetes, and apicomplexans. Phototrophs were mostly chrysophytes, chlorophytes, and dinoflagellates. Diatoms were less abundant compared to typical assemblages in lakes and rivers. Only 16 unique diatom sequences were obtained, which indicates that diatoms were less abundant in ponds compared to other microbial organisms and that a diatom-specific *rbc*L sequencing is necessary to reveal their diversity in this environmental setting. The diatom genera for which the sequences were obtained and taxonomically assigned were *Eunotia, Neidium, Navicula, Pinnularia*, *Stenopterobia*, and *Ulnaria*.

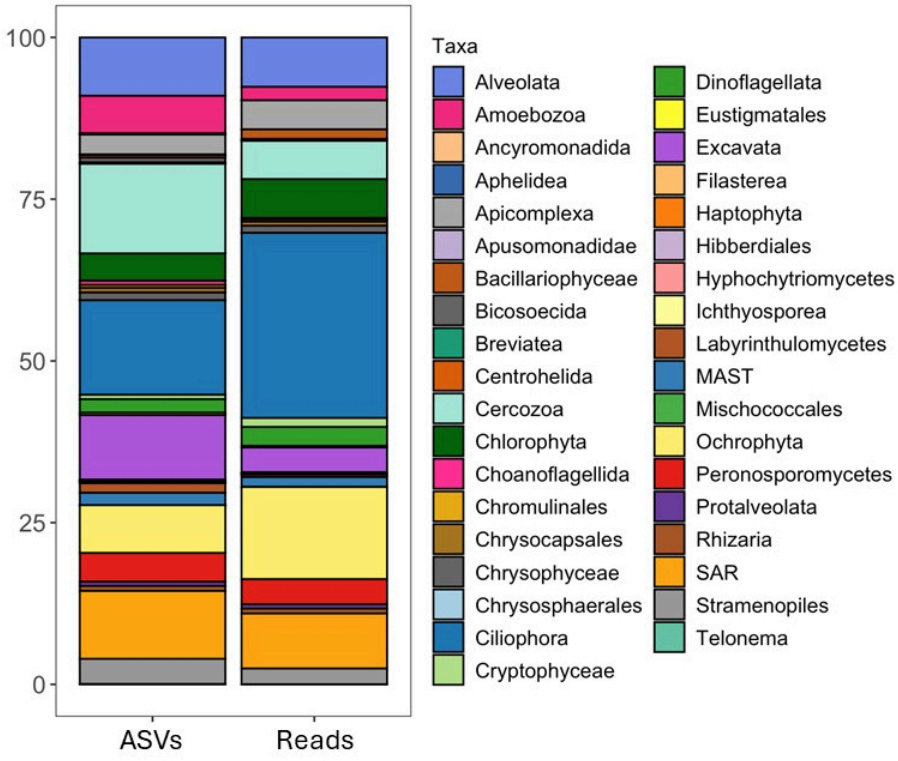

Figure 2. Percentage of 18S sequences obtained with unique sequence amplicons (ASVs) - left panel, and reads - right panel, assigned to major taxonomic groups of protists.

The NMDS ordination shows that protists communities are strongly impacted by water chemistry, including nutrient and potentially road salt pollution estimated by chloride concentrations (Fig. 3). It also shows significant differences in composition of protistan assemblages inhabiting different microhabitats. Various taxa, such as, for example, peritrichian ciliates *Epistylis hentschelia* and *Vorticella* sp. were associated with more acidic and low-nutrient ponds, while several algal taxa, such as green alga *Spirogyra* sp. and dinoflagellates *Woloszynskia pascheri* and *Cystodinium bataviense* were more common in relatively polluted ponds.

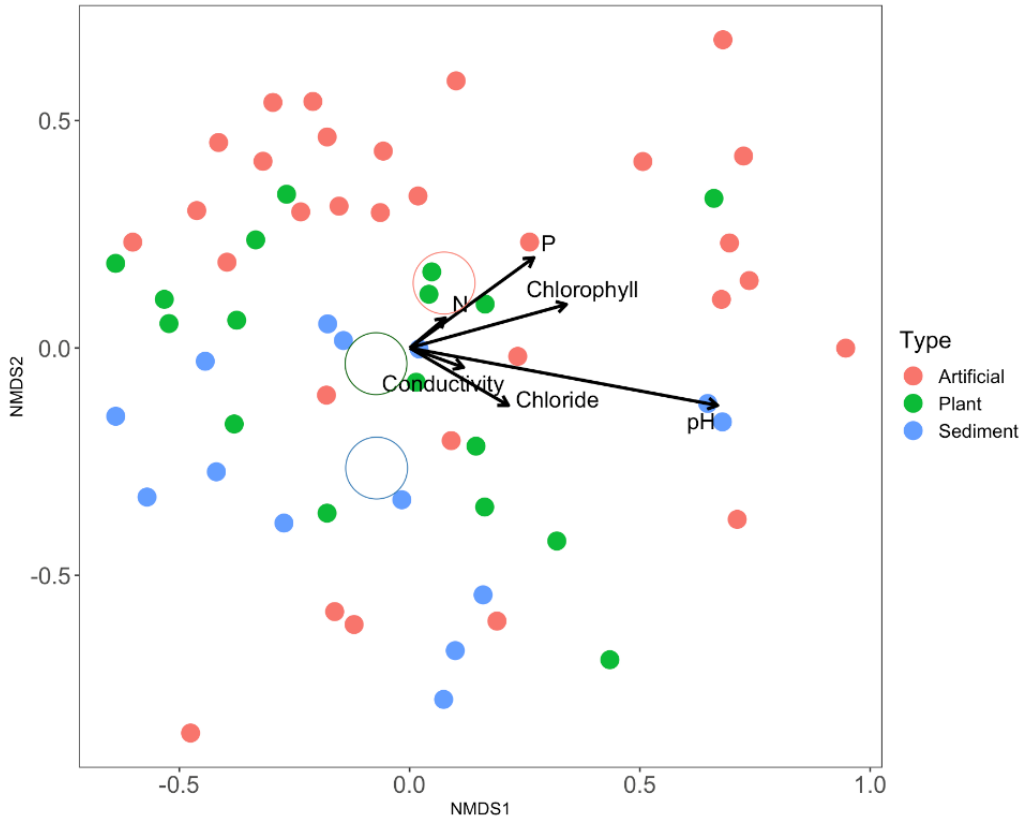

Figure 3. NMDS ordination of 18S protist dataset showing position of 61 samples collected from diatometer slides (red circles), aquatic plants (green circles) and pond surface sediments (blue circles). Empty circles show centroids of these three types of substrates. The arrows show the correlations of water quality characteristics with ordination axes.

*3.2. Diatom rbc*L *metabarcoding*

*Rbc*L sequencing yielded a total of 2,069,069 raw reads. 1,044,706 reads remained after quality filtering, merging, denoising and chimera removal. Out of 2,991 ASVs represented by these reads, 1,320 were taxonomically assigned to diatoms (Bacillariophyta), with 370 diatom ASVs remaining in the dataset after removing sequences not assigned below the phylum level (Appendix 2).

319 or 86% of diatom ASVs were taxonomically assigned to genus level using the Diat.barcode reference database; these constituted 98.7% of all diatom reads. The genera most represented in terms of ASV numbers were *Pinnularia* (72 ASVs identified to genus level), *Eunotia* (65), *Nitzschia* (41) and *Neidium* (14). The genera with the highest abundance of reads were *Eunotia* (31% of all reads), *Pinnularia* (26%), *Nitzschia* (12%), *Gomphonema* (10%),

Frustulia (7%), and *Achnanthidium* (3%). The differences among the tree types of substrates in terms of ASV diversity and reads proportions is shown in Fig. 4.

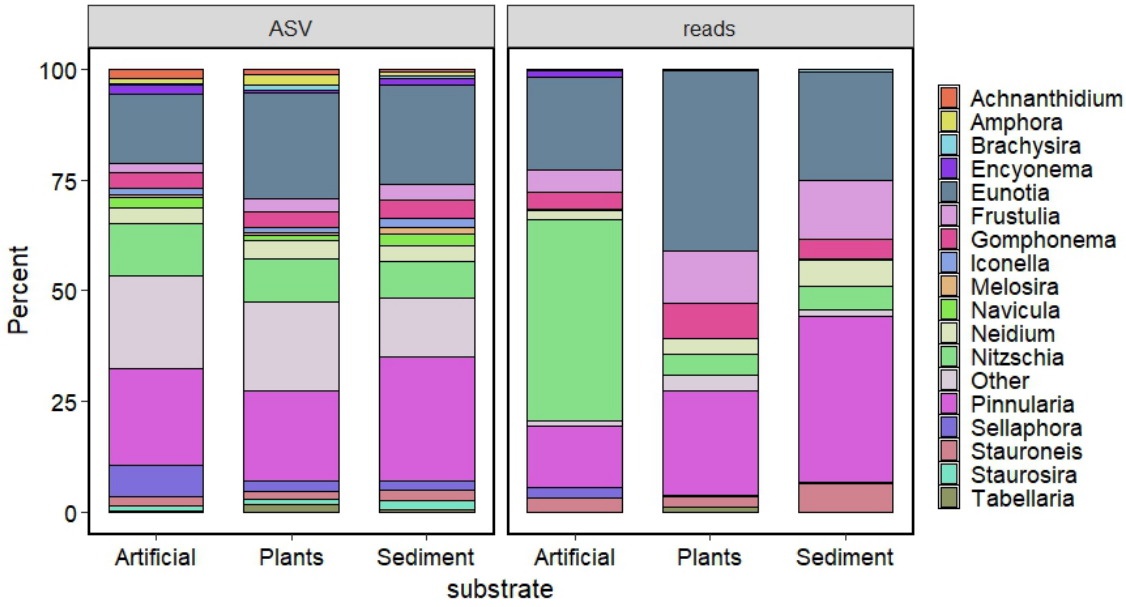

Figure 4. Percentage of *rbc*L sequences obtained with ASVs (left panel) and reads (right panel) assigned to various diatom genera collected from multiple substrates.

Out of 370 ASVs, 197 or 53% were identified to species level with at least 75% accuracy, while only 76 ASVs (21%) were identified to species with 100% accuracy. The reads that belonged to the ASVs identified to species at 100% level constitute 18% of the total number of reads in the *rbc*L dataset. This indicates that the diatoms common in studied ponds are poorly covered by the Diat.barcode.

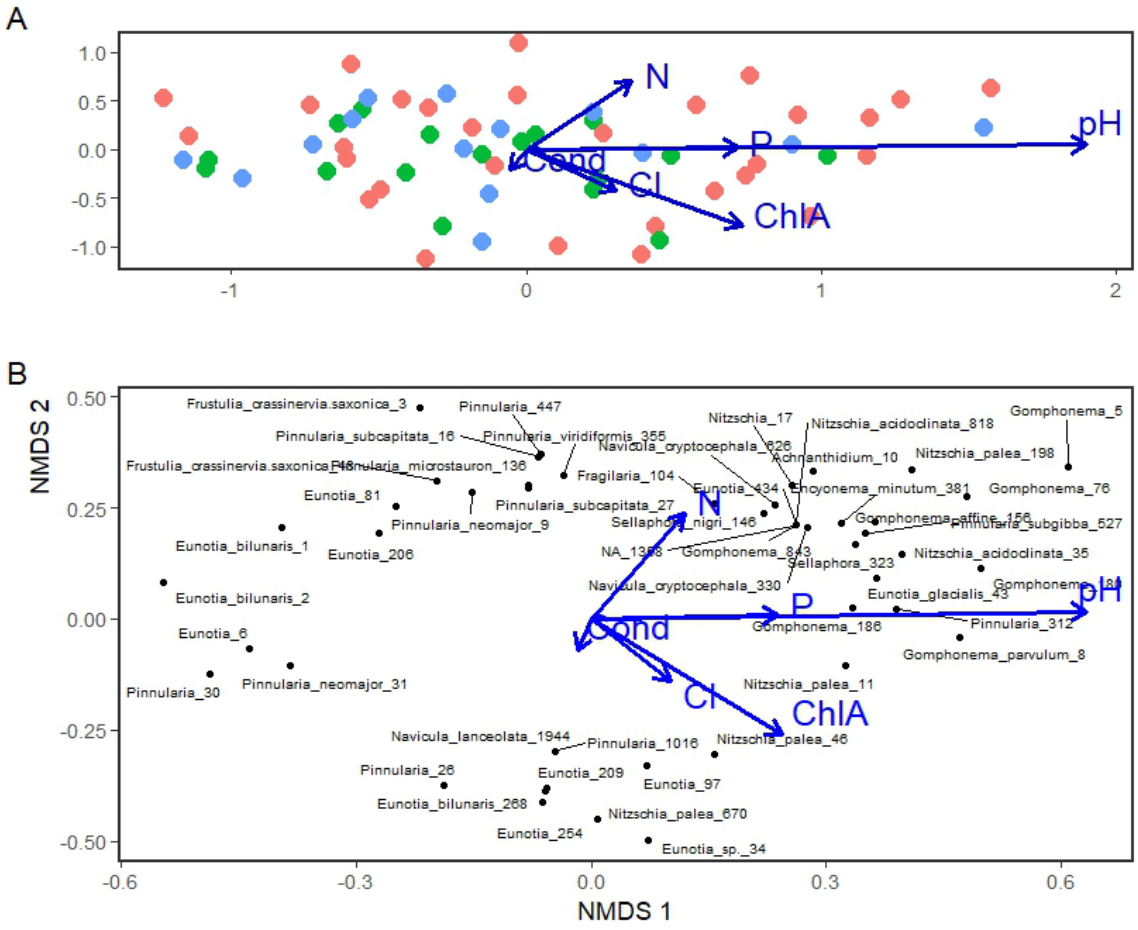

Figure 5. NMDS ordinations of *rbc*L dataset showing (A) positions of samples from diatometer slides (red), plants (green) and sediments (blue) and (B) positions of ASVs with the greatest fit ($p<0.05$) to the ordination. The arrows show correlations of measured water-quality parameters with ordination axes.

NMDS ordination coupled with *envfit* procedure (Fig. 5) showed that diatoms, as characterized by *rbc*L barcodes, were most sensitive to pH variation. At the same time, positions of ASVs indicate diatom assemblage response to nutrients too, as the taxa typically associated with eutrophication (*Navicula, Nitzschia, Sellaphora*) are concentrated in the lower right quadrant of the ordination space.

Mantel tests assessed the strength and significance of the relationships between *rbc*L diatom data and measured environmental parameters (Table 2). The tests used Spearman rank correlation between *rbc*L diatom matrix and the matrix composed of environmental parameters and compared responses of assemblages collected from three types of substrates in 13 ponds. The strongest association between assemblage composition and the set of available environmental parameters as

estimated by the Mantel r statistic was found for sediment samples, while diatometer samples did not show any relation to the environment. This must be interpreted as the result of the considerable time lag between water quality and diatom sampling, which was on the order of several years for some samples. Diatometer samples represent pioneer assemblages formed in two weeks and should reflect water quality in this short time period only. They also may be influenced by founder effects with first colonizers being a relatively random subset of species inhabiting the whole ecosystem.

Table 2. Results of the Mantel test assessing the strength and significance of the relationships between *rbc*L diatom data and measured environmental parameters.

Dataset	**R statistic**	**Significance**
All 61 samples	0.204	0.0002
42 samples from 13 ponds where both natural and artificial substrates were sampled	0.185	0.0031
13 sediment samples from 13 ponds	0.253	0.0605
16 plant samples from 13 ponds	0.145	0.1388
13 diatometer samples from 13 ponds	-0.530	0.6062

3.3 Diatom counts

A total of 247 diatom taxa were recorded microscopically in 61 pond samples (Appendices 3 and 4). Identification of several taxa to species level was uncertain as specimens morphologically graded from one species to another. This was especially true for most abundant taxa in the genera *Eunotia*, *Frustulia* and *Gomphonema*. The most common representatives of *Eunotia* were long-celled species that could have been identified as *Eunotia naegeli, E. juttneri* or *E. genuflexa*. These species differ in their valve outline (Lange-Bertalot et al. 2011) and in studied ponds it was impossible to separate them as specimens withing and among populations seemed to vary continuously in shape. Another species complex is *Gomphonema graciledictum/ G. naviculoides/G. subnaviculoides*. Likewise, the species in *Frustulia saxonica/F. crassinervia* complex morphologically grade into each other and exhibit a considerable variation within and between populations.

NMDS ordination of all 61 samples revealed the major gradient in assemblage composition related to water pH and the second gradient associated with nutrient enrichment (Fig. 6). The ponds with higher phosphorus and chlorophyll A concentrations had a higher proportion of species typically found in inland waterbodies with moderate to high nutrient content, such as *Navicula* spp., *Nitzschia* spp., *Sellaphora* spp., and *Achnanthidium* spp. The low-nutrient ponds had a higher

proportion of *Eunotia, Pinnularia* and *Gomphonema* with *Eunotia* dominating acidic ponds and *Gomphonema* being more abundant in higher-pH waters.

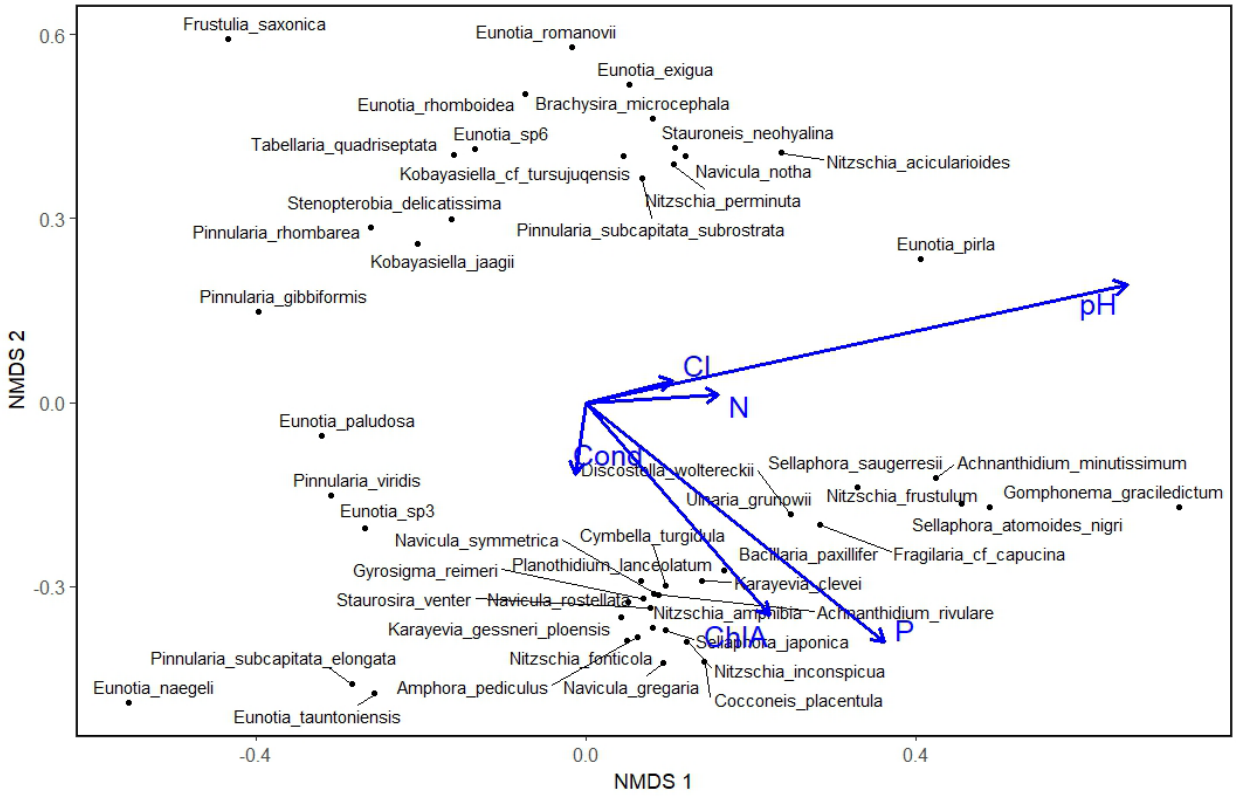

Figure 6. NMDS ordination of diatom count dataset showing position of 49 species with the greatest fit ($p<0.05$) to the ordination. The arrows show significant ($p<0.05$) correlations of water quality characteristics with ordination axes.

3.4 Diatom diversity estimated by microscopy and metabarcoding

Establishing correspondence between morphospecies and DNA sequences from environmental metabarcoding depends on the availability of low-diversity samples. Therefore, both taxa richness and evenness were estimated in studied samples. As the diatom counting aimed at reaching the 400-valves threshold, the same number was used for rarefying the *rbc*L data matrix. Comparisons of taxa/ ASV richness between diatom count and *rbc*L data (Fig. 7A) indicates that only a few samples, mostly from diatometer slides, had very low number of species. Shannon diversity indices followed the same pattern (Fig. 7B) with only a few samples having the index value below 1 for both morphotaxa and *rbc*L datasets.

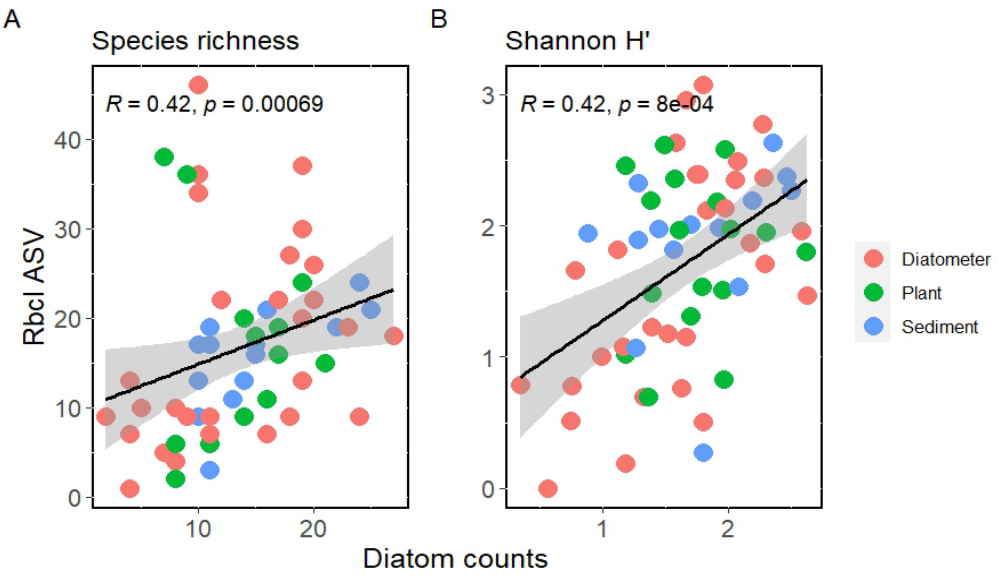

Figure 7. Species vs ASV richness (A) and Shannon diversity (B) in samples from different substrates.

3.5. Diatom barcodes

The following sections describe the most successful attempts at establishing correspondence between morphotaxa and *rbc*L barcode sequences.

3.5.1. Centric diatoms

Centric diatoms were not abundant in sampled ponds, but a few morphospecies of *Aulacoseira, Cyclotella, Discostella, Melosira varians* and *Thalassiosira weissflogii* were encountered in the counts. The co-occurrence analysis revealed a single association between the presence of *Discostella woltereckii* and ASV 518 that could be matched to the level of *Discostella* genus using the Diat.barcode database and found only in a sample from Windsor Basin (NJPC43). This sequence is 100% identical to the Genbank sequence OQ504857.1 identified as *Discostella* sp. clade L strain QM3 from Quebec, Canada. This clade of *Discostella* could not be linked to known morphospecies according to Schultz et al (2024) and may represent a new species morphologically resembling *D. woltereckii*.

3.5.2. Araphid diatoms

There was only one *Ulnaria* and two *Fragilaria* ASVs in the *rbc*L dataset and their relative abundance neatly matched presence of these genera in the microscope counts.

Ulnaria grunowii

A single sample NJPC43 had a relatively high abundance (4.5%) of a large-celled diatom that morphologically conformed to the description of *Ulnaria grunowii* (Lange-Bertalot & S. Ulrich) Cantonati & Lange-Bertalot (Fig. 8A). The *rbc*L sequence that unequivocally corresponded to this morphospecies (ASV42, 14.1% of reads) 100% matched the Genbank accession HQ912454.1, which is the *rbc*L sequence of the strain UTEX FD404, identified by the UTEX collection as *Synedra ulna* var. *chaseana* B.W. Thomas, with the name shortened to "*Synedra ulna*" in the Genbank. The isotype material of *S. ulna* var. *chaseana* housed at ANS (Fig. 8B) shows a diatom 3-5 um wide, 400-700 μm long with 7-11 striae in 10 μm with capitate ends and without distinct central area, which corresponds to the original description (Walker & Chase 1886). It is most likely that the UTEX strain FD404 was identified as *S. ulna* var. *chaseana* based on the illustrations of that taxon in Patrick & Reimer (1966), but representing the same species as found in our sample from Windsor Basin (NJPC43), with shorter valves and considerably higher striae density (13-15 in 10 μm), indicating its placement in *Ulnaria grunowii*.

The recently published 18S rDNA phylogeny of needle-shaped *Ulnaria* and *Fragilaria* taxa (Zakharova et al. 2023) placed the UTEX strain FD404 in a clade (Uln1) that contains several strains of relatively slender *Ulnaria* that the authors interpreted as a single species *Ulnaria acus*. This opinion, is however, likely to be refuted by those who relatively recently described numerous species that would morphologically fit into this group (Lange-Bertalot and Ulrich 2014, Kulikovsky et al. 2016, Alexon et al. 2020). Interestingly, the Diat.barcode assigns ASV42 with 100% certainty to *Ulnaria ulna*, which indicates that these assignments should not be taken at face value and may need to be critically revised.

Figure 8. Selected araphid diatoms. A- *Ulnaria grunowii*, Windsor Basin, matching ASV42, B – "*Synedra ulna* var. *chaseana*", ANSP HLSmith588, isotype material, C – *Fragilaria* cf. *capucina*, Windsor Basin, matching ASV104.

Fragilaria cf. *capucina*

Among the three *Fragilaria* taxa recorded in the counts, only *Fragilaria* cf. *capucina* (Fig. 8C) reached considerable abundance in two samples, 22.3% in Windsor Pond (NJPC43) and 3% in Country Pond (NJPC46). These same samples were the only ones with the presence of ASV104 at 6.1% and 1% of reads correspondingly (besides only one other sample where it was at very low abundance (0.02%). This ASV was a 100% match to several Genbank *rcb*L sequences representing *Fragilaria* strains isolated from Eurasia and either not identified to species level (BZ15, BZ35, Lab56, 032FraR02) or identified as *F. crotonensis* (TCC301), *Centronella reicheltii* (CCAP1011), and *Fragilaria bidens* (s0327) and forming a clade with other *Fragilaria* species morphologically similar to *F. crotonensis, F. perminuta* and other species typically capable of making ribbon-like colonies (Zakharova et al. 2023). The gap analysis carried out by Zakharova et al. (2023) suggested that these strains represent a single species, with a proper name to be determined upon further studies.

Small fragilariod diatoms identified microscopically as *Staurosira* or *Pseudostaurosira* spp. could not be matched with certainty to specific ASVs as they were not especially abundant in studied samples and there was a diversity of both morphologies and sequences in samples where these organisms occurred.

Tabellaria quadriseptata

While there were several morphological entities of *Tabellaria* recorded in the microscope counts, only three ASVs were identified as *Tabellaria*, all as *T. flocculosa*. ASV 253 and 303 were assigned to *T. flocculosa* with 100% certainty, while ASV 188 was assigned to that species at 84% level. Co-occurrence analysis suggested that ASV 188 is a strong match to *Tabellaria quadriseptata*, while the other two ASVs could not be matched clearly to one or another informal "Koppen" groups of *T. flocculosa* recognized by US diatom analysts.

3.5.3. The genus Eunotia

The distance tree of *Eunotia* barcode sequences (Fig. 9) shows that ASVs found in New Jersey Pineland ponds were scattered almost across the whole tree and some formed clusters that did not include any sequences from Diat.barcode database. This is an indication of taxa that might be limited in their distribution or, alternatively, the taxa that have not been sequenced yet.
Three clusters of ASVs were matched to morphological entities (Fig. 9).

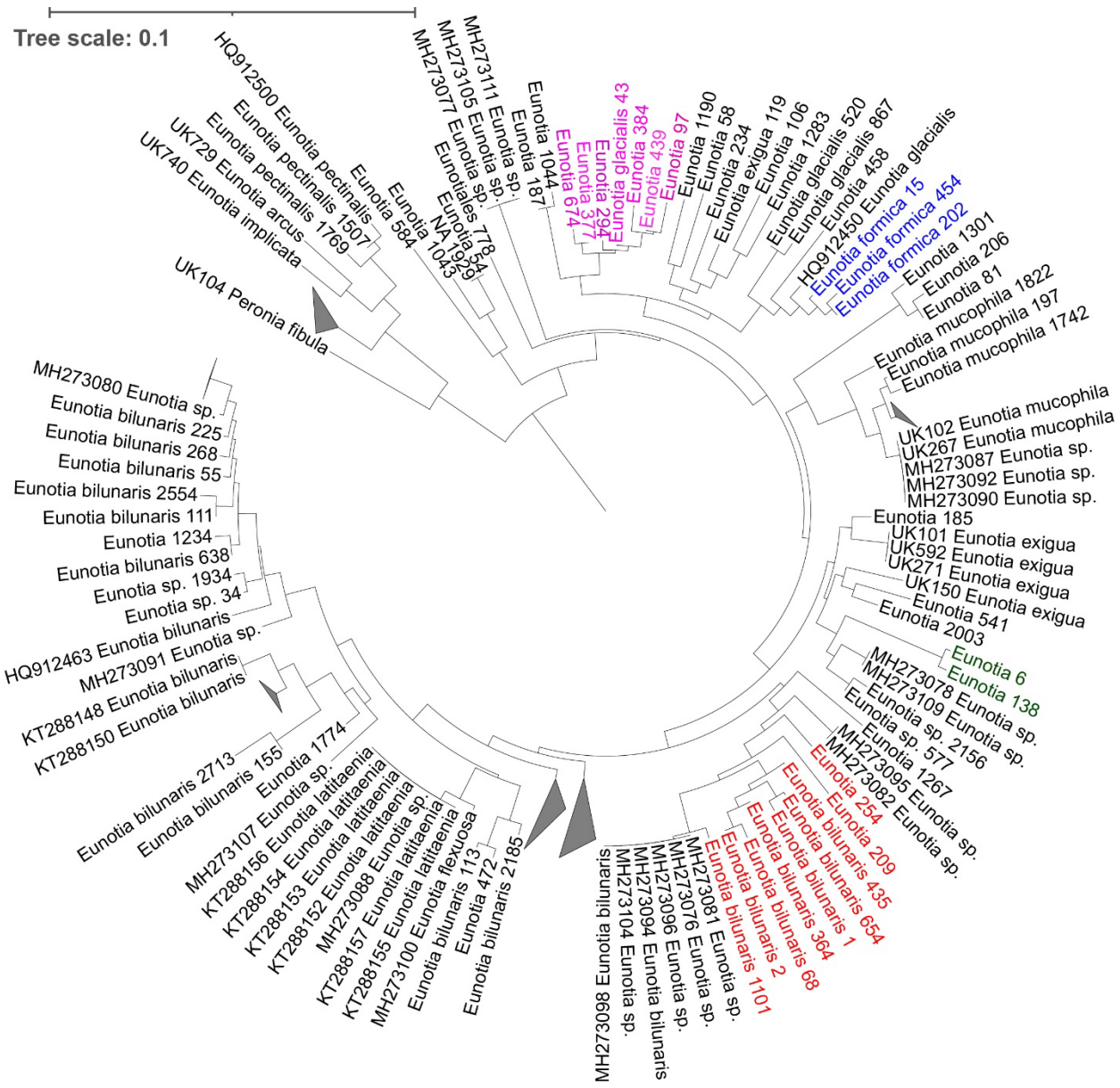

Figure 9. A portion of the Neighbor-joining tree showing positions of *rbc*L sequences that belong to the genus *Eunotia*. The names of ASVs encountered in the studied samples start with the word Eunotia and end with ASV number, while the names of sequences from the Diat. Barcode database begin with the Genbank accession number and end with the Latin species name. Several clades not containing sequences found in studied ponds are collapsed. ASVs shown in red represent *E. genuflexa/naegeli* species complex; purple – *E. pirla*; blue – *E. metamonodon*; green – *E. tauntoniensis.*

Eunotia naegeli/genuflexa species complex

The sample from Pennypot Pond (NJPC44) was 100% dominated by the *Eunotia naegeli/genuflexa* species complex (Fig. 10). The two most abundant reads belonged to ASV1:

80.8%; ASV2: 17.4%. Both sequences were assigned to "*Eunotia bilunaris*" by Diat.barcode and had the closest blast matches (7 substitutions or 97% identity) to several identical Genbank sequences from the following isolates *Eunotia* sp. voucher HELL19 (MH273104.1); *Eunotia bilunaris* voucher SPAG24 (MH273098.1); *Eunotia* sp. voucher KALM1_5 (MH273096.1); *Eunotia bilunaris* voucher NOS14 (MH273094.1), *Eunotia* sp. voucher KALM1_3 (MH273081.1), *Eunotia* sp. voucher JL1-1(MH273076.1), all from European collections. The ASV1 and ASV2 differed from each other by two base pairs and could be safely considered to represent the *Eunotia naegeli/genuflexa* species complex:

```
ASV2  1    CGTTACTGCAGCTACTCAAGAAGAAGTTTACAAACGTTCAGAATTCGCTAAAGAACTTGG  60
           ||||||||||||||||||||||||||||||||||||| || ||||||||||||||||||||
ASV1  1    CGTTACTGCAGCTACTCAAGAAGAAGTTTACAAACGTGCACAATTCGCTAAAGAACTTGG  60

ASV2  61   TTCTGTAATTATTATGATCGACTTAGTAATGGGCTATACAGCAATCCAAACAATTGCTCT  120
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ASV1  61   TTCTGTAATTATTATGATCGACTTAGTAATGGGCTATACAGCAATCCAAACAATTGCTCT  120

ASV2  121  TTGGGCTCGTGAACACGATATGATTTTACATTTACATCGTGCAGGTAACTCTGCATATGC  180
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ASV1  121  TTGGGCTCGTGAACACGATATGATTTTACATTTACATCGTGCAGGTAACTCTGCATATGC  180

ASV2  181  TCGTCAAAAAAATCATGGTATTAACTTCCGTGTTATTTGTAAATGGATGCGTATGTCTGG  240
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ASV1  181  TCGTCAAAAAAATCATGGTATTAACTTCCGTGTTATTTGTAAATGGATGCGTATGTCTGG  240

ASV2  241  TGTAGATCATATCCATGCTGGTA  263
           ||||||||||||||||||||||
ASV1  241  TGTAGATCATATCCATGCTGGTA  263
```

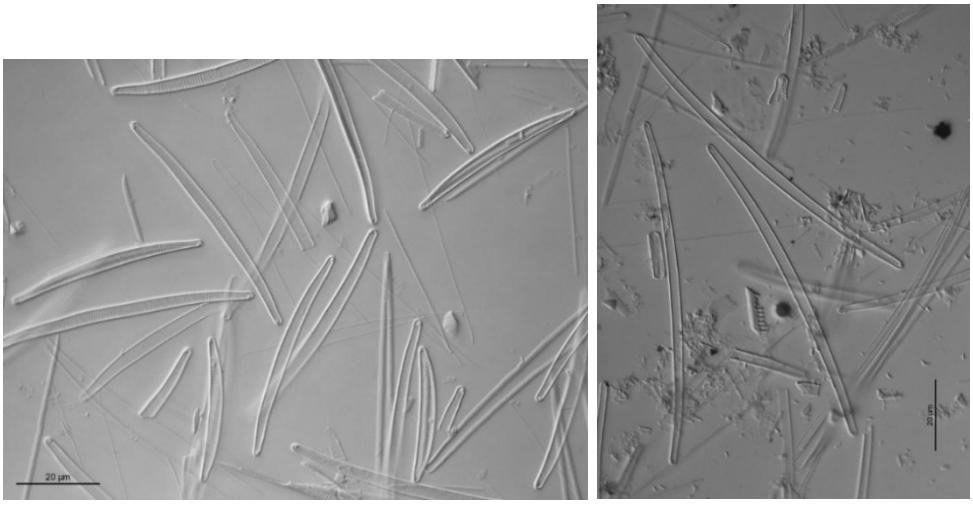

Figure 10. Samples dominated by *E. naegeli/genuflexa* species complex. Left: Pennypot Pond, sample NJPC44, right: Basket Pond, NJPC50.

It should be noted that in the original counts an attempt was made to separate *E. genuflexa* from *E. naegeli*, but there was not correspondence between these two entities and the occurrence of ASV1 and 2 and the separation was quite arbitrary. The distance analysis of all *Eunotia* sequences revealed a cluster of ASVs that besides ASV1 and ASV2 contains ASVs #68, 209, 254,

364, 435, 654, and 1101. All these sequences likely represent the same *E. naegeli/genuflexa* species complex characteristic for Pinelands.

Five other samples that had relatively high counts of this species complex and ASV1 or ASV2 or ASV68 were present in all these samples (Table 3). In sample NJPC59 only the ASV2 was present at 6.2% relative abundance, while two other *Eunotia* sequences (ASV54 and ASV68) had a higher percentage. ASV54 may represent *E. lewisii*, which has much higher cell volume than *E. naegeli/genuflexa* and was identified morphologically in this sample. The relatively low abundance of ASV1 in the sediment sample from Price Pond (PricSedm23) can be explained by the considerable abundance (23% of diatom count) of a high cell-volume diatom *Frustulia saxonica*, which also constituted 44% of *rbc*L reads count.

Table 3. Percentage of *Eunotia naegeli/genuflexa* in microscope counts and ASV1, 2, and 68 reads in the *rbcL* data matrix.

Pond/Sample	Count of *Eunotia naegeli/genuflexa*	ASV1	ASV2	ASV68
Pennypot/NJPC44	100	80.8	17.4	-
Basket/NJPC50	79	23.7	38.1	-
Wesickaman/WesiSlid23	79	-	89.2	-
Forbidden/NJPC59	63	-	6.2	25.3
Holly/HollPlnt23	52	6.1	32.7	-
Price/PricSedm23	51	3.8	-	-

Eunotia pirla

Two samples had a high valve count of *Eunotia pirla* (Fig. 11), but the *rbc*L reads had two different dominant *Eunotia* ASVs: ASV 43 and ASV97 with 3 bp difference:

```
ASV97  1    CGTTACAGCAGCTACTCAAGAAGAAGTTTACAAACGTGCAGAGTTTGCTAAAGAACTTGG  60
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ASV43  1    CGTTACAGCAGCTACTCAAGAAGAAGTTTACAAACGTGCAGAGTTTGCTAAAGAACTTGG  60

ASV97  61   TTCTGTAATTGTTATGATAGACTTAGTAATGGGTTACACCTCAATTCAAACAACTGCTAT  120
           |||||||||||||||||| |||||||||||||||||| |||||||||||||||||||||
ASV43  61   TTCTGTAATTGTTATGATCGACTTAGTAATGGGTTACACATCAATTCAAACAACTGCTAT  120

ASV97  121  TTGGGCACGTGAAAATGATATGATTTTACACTTACACCGTGCAGGTAACTCTACATATGC  180
           ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ASV43  121  TTGGGCACGTGAAAATGATATGATTTTACACTTACACCGTGCAGGTAACTCTACATATGC  180

ASV97  181  TCGTCaaaaaaaTCATGGGATTAACTTCCGTGTTATTTGTAAATGGATGCGTATGTCTGG  240
           |||||||||||||||||| ||||||||||||||||||||||||||||||||||||||||
ASV43  181  TCGTCAAAAAAATCATGGTATTAACTTCCGTGTTATTTGTAAATGGATGCGTATGTCTGG  240

ASV97  241  TGTAGACCATATCCACGCTGGTA  263
           |||||||||||||||||||||||
ASV43  241  TGTAGACCATATCCACGCTGGTA  263
```

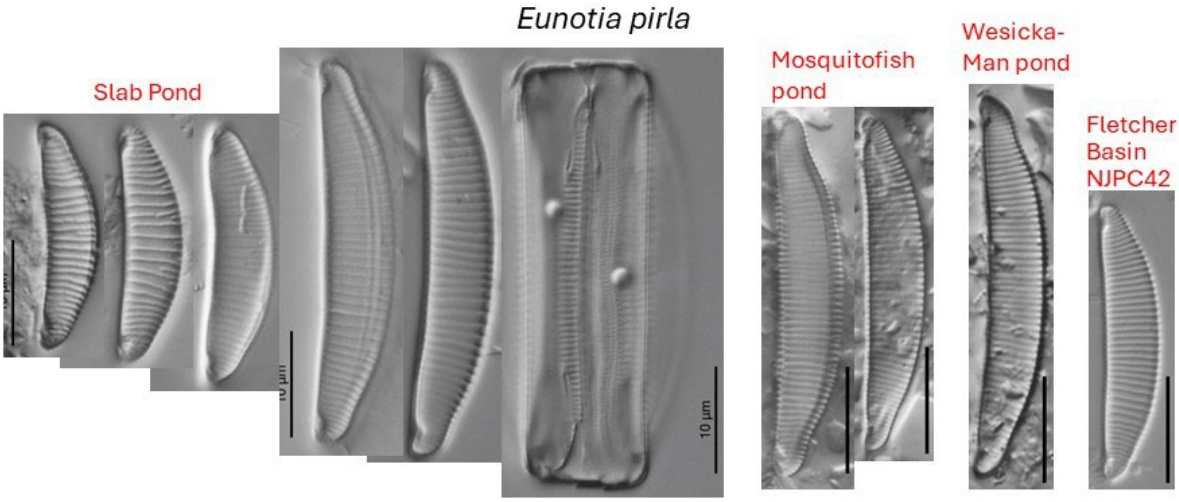

Figure 11. *Eunotia pirla* in New Jersey Pineland ponds.

A sediment sample from Slab Pond (SlabSedm22-1) had 93% *E. pirla* in the count and corresponding ASV43 at 65% relative abundance. The vegetation sample from Mosquitofish Pond (MosqPlnt22-1) had 80.6% *E. pirla* in the valve count and 37.6% of ASV97 in the *rbc*L data. When matched to Genbank sequences, both ASVs had the closest match to the sequence HQ912450.1 identified as *Eunotia glacialis* (8-9 bp difference) and to the sequence AM710428.1 identified as *E. formica* (9-11 bp difference). The cluster of sequences similar to ASV43 and 97 also contained ASV294, 377, 439, and 674. These ASVs were found in the following ponds: Cooper, Evans, Flittertown, Island, Mosquitofish, Fletcher, Windsor, and Slab. *E. pirla* was present in diatom counts from Cooper, Holly, Island, Mosquitofish, Fletcher, Leah, Price, Slab, and Wesickaman ponds.

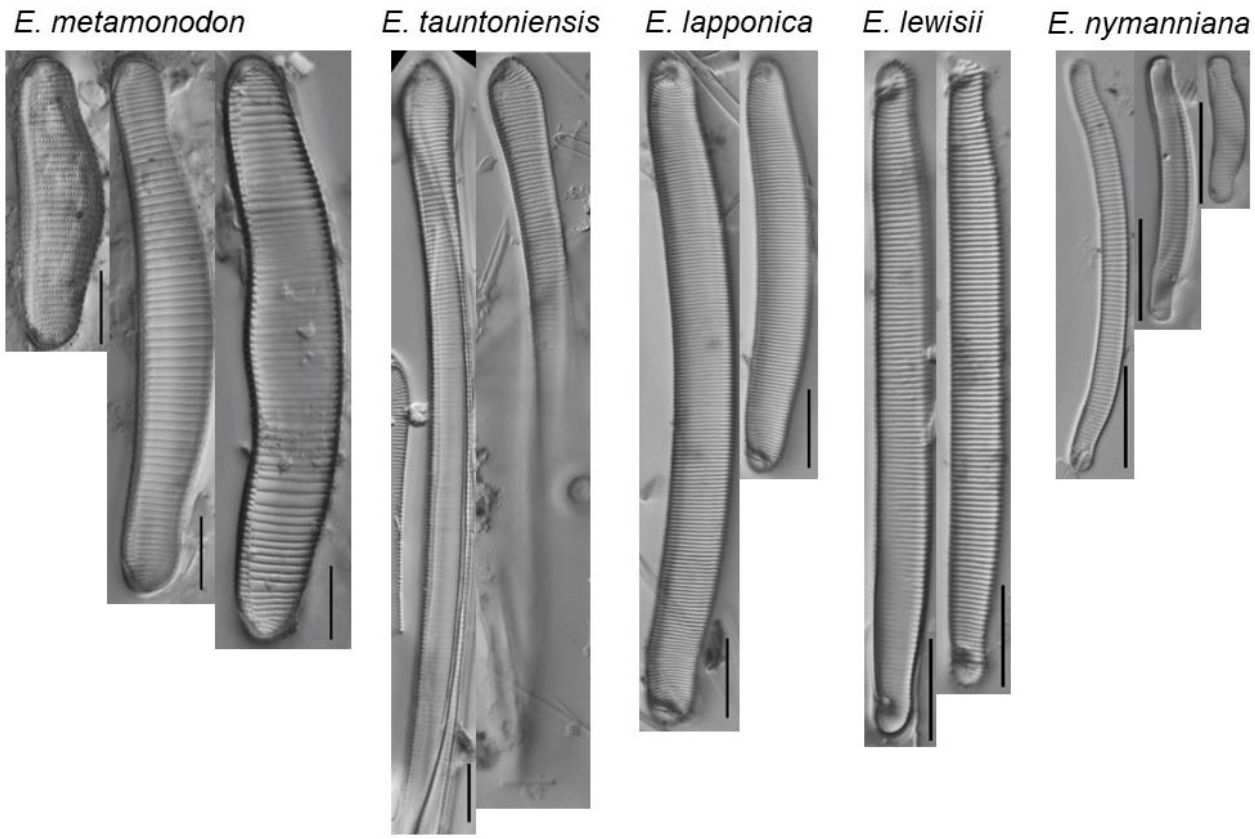

Figure 12. Selected *Eunotia* morphospecies from the New Jersey Pinelands ponds.

Eunotia metamonodon

This large-celled species (Fig. 12) was present only in Cooper Pond and reached a relatively high abundance (22% of the count) in the aquatic vegetation sample. The same sample was dominated by three similar amplicons (ASV15, 68%, ASV202, 7%, and ASV454, 2%) that were exclusively found in this sample and were assigned by Diat.barcode to *Eunotia formica*. These ASVs differed by one base pair from each other and most closely matched the sequence AM710428.1 identified as "*Eunotia formica*" by Bruder & Medlin (2007), from which they differed by one or and two base pairs. *E. metamonodon* is morphologically similar to *E. formica* and these species may represent a complex of closely related lineages. Moreover, as Bruder & Medlin Genbank sequences were not documented by voucher specimens of photographs, it is impossible to exclude a possibility that their "*E. formica*" was indeed *E. metamonodon* or another similar species.

```
ASV15  1    CGTTACTGCAGCTACTCAAGAAGAAGTTTACAAACGTGCTGCTTTTGCTAAAGAACTTGG  60
            ||||||||| ||||||||||||||||||||||||||||||||||||||||||||||||||
ASV202 1    CGTTACTGCCGCTACTCAAGAAGAAGTTTACAAACGTGCTGCTTTTGCTAAAGAACTTGG  60

ASV15  61   TTCTGTTATTATTATGATCGACTTAGTAATGGGTTACACATCAATCCAAACAACTGCTAT  120
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ASV202 61   TTCTGTTATTATTATGATCGACTTAGTAATGGGTTACACATCAATCCAAACAACTGCTAT  120

ASV15  121  TTGGGCACGTGAAAACGATATGATTTTACATTTACACCGTGCAGGTAACTCTACATATGC  180
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ASV202 121  TTGGGCACGTGAAAACGATATGATTTTACATTTACACCGTGCAGGTAACTCTACATATGC  180

ASV15  181  TCGTCaaaaaaaTCATGGTATTAACTTCCGTGTTATTTGTAAATGGATGCGTATGTCTGG  240
            ||||||||||||||||||||||||||||||||||||||||||||||||||||||||||||
ASV202 181  TCGTCAAAAAAATCATGGTATTAACTTCCGTGTTATTTGTAAATGGATGCGTATGTCTGG  240

ASV15  241  TGTAGACCATATCCACGCTGGTA  263
            |||||||||||||||||||||||
ASV202 241  TGTAGACCATATCCACGCTGGTA  263
```

The three other sequences clustered together with ASV15 and ASV202 were ASV458, 520 and 867. ASV 520 and 867 were also found exclusively in the same sample and may represent the same lineage morphologically corresponding to *E. metamonodon*.

Eunotia tauntoniensis

The co-occurrence analysis identified two sequences, ASV 6 and 138 as the most strongly associated with *Eunotia tauntoniensis*, a large-celled diatom that was quite abundant in several samples, mostly from Flittertown, Holly, and Wesickaman Ponds (Fig. 12). These two ASVs clustered together in the neighbor-joining tree (Fig. 9) and appear as a sister group to several *Eunotia* isolated from Europe, not identified to species in Diat.barcode. The ASV6 makes a large proportion of reads in samples from the above-mentioned ponds, which is attributed to the large size of *E. tauntoniensis*. ASV 6 is 100% identical in its barcode region to the GenBank sequence KM999115.1 (Eunotia sp. KEL-2015 clone JAR78 from the Big Moose Lake, New York, USA), while ASV 138 has 1 bp difference with it.

Eunotia nymanniana

The co-occurrence analysis indicated a correspondence between *Eunotia nymanniana* (Fig. 12) and ASV 185, both found only in two samples from Evans Pond. The ASV 185 could not be identified to species level with Diat.barcode and did not closely match any GenBank *Eunotia* sequence, but it is most similar to the *E. exigua* clade (Fig. 8), which makes sense as *E. nymanniana* is somewhat morphologically similar to *E. exigua.*

Eunotia lapponica (Fig. 12) may be represented by ASVs 81, 206, and 1301, but establishing a strong correlation requires further study.

3.5.4. The genus Gomphonema

While representatives of *Gomphonema* were often dominant in both diatom counts and in *rbc*L data, only ten *Gomphonema* ASVs were recorded. In comparison, six *Gomphonema* morphospecies were recorded in these samples by the author of this report and 16 by Dr. M. Enache. This discordance underscores the difficulty of separating morphological entities within this genus using subtle differences in diatom valve shape, size and stria pattern.

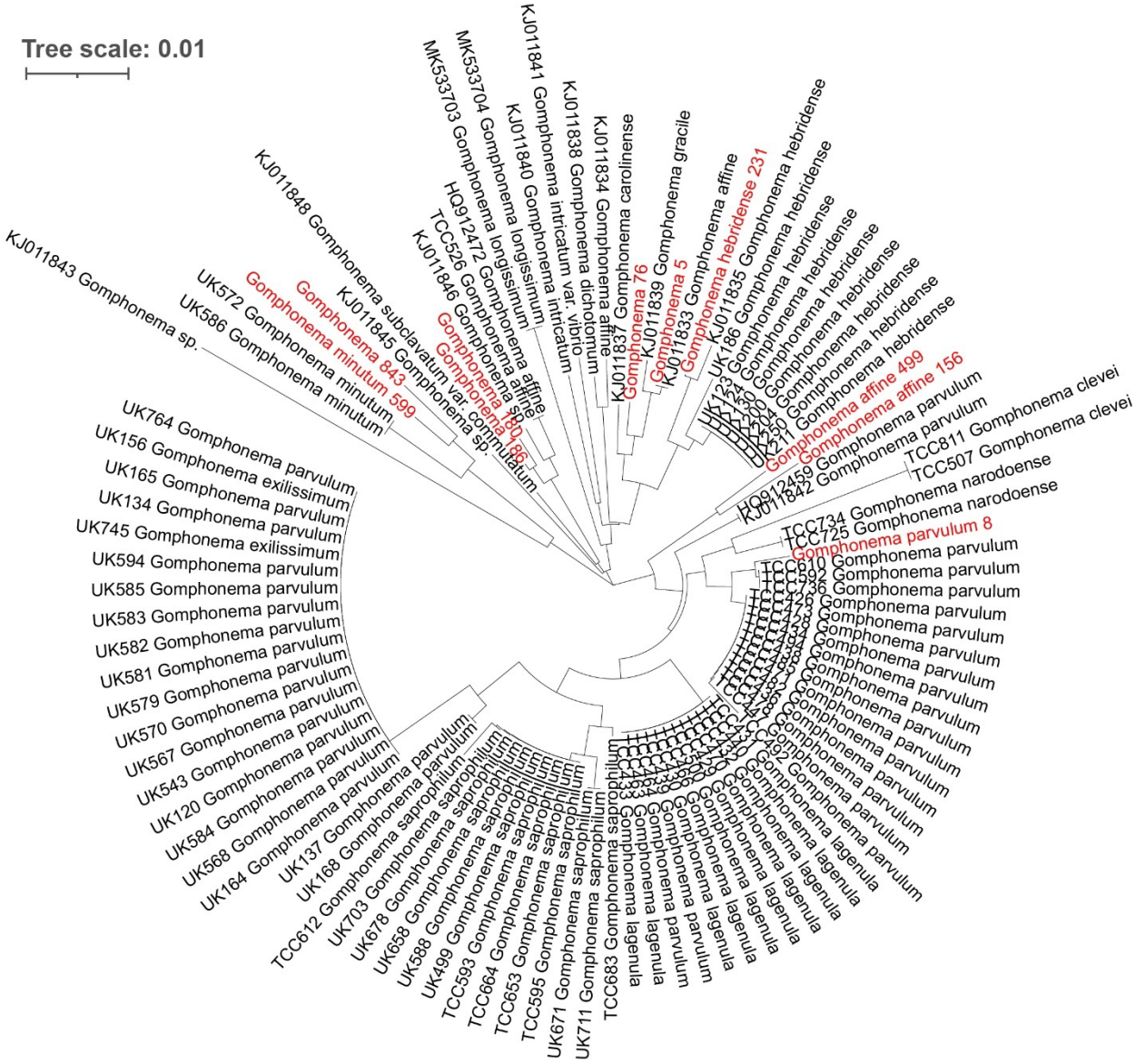

Figure 13. A portion of the Neighbor-joining tree showing positions of rbcL sequences that belong to the *Gomphonema* clade. ASV encountered in the studied samples are shown in red; the rest of sequences are from the rbcL alignment in Diat.barcode database v.10.1.

Fig. 13 shows the placement of the 10 *Gomphonema* ASVs found in studied samples within a portion of the neighbor-joined tree containing most similar *Gomphonema* sequences. These sequences are found in five clusters; with the most abundant sequences found in cluster of ASV 5+76+231 and the second most abundant cluster consisting of ASV 8. The first cluster (ASV 5+76+231) contains sequences identified in the Diat.barcode database as *G. carolinense*, *G. gracile, G. affine* and *G. hebridense*, while the second (ASV 8) contains only *G. parvulum*.

The morphospecies of *Gomphonema* with the overall highest abundance in diatom counts is named here "*G. graciledictum/subnaviculoides"*. Dr. Enache split them into several groups referring to the largest specimens as *Gomphonema* aff. *affinis* while placing others in *Gomphonema* sp. 3 and sp. 4. The author of this report could not draw consistent boundaries between these entities and lumped them into a species complex. Morphometrically, the specimens observed in the New Jersey ponds better fit into *G. subnaviculoides* than into other species with similar morphology, but they are considered to represent a species complex because (1) there is an overlap in all major morphological characteristics between species constituting this group (Table 4) and (2) because ecologically they better fit to the description of *G. graciledictum* than *G. subnaviculoides.* The Pinelands ponds where these specimens reached at least 50% of relative abundance in at least one sample were slightly acidic (pH 5-6.6, mean 5.7) and of low mineral content (conductivity 45-295, mean 76 μS/cm), while *G. subnaviculoides* has been described from a meso-to eutrophic lake Prespa with moderately hard water (Levkov et al. 2016).

Table 4. *Gomphonema* morphometrics. "*G. graciledictum/ subnaviculoides* NJ" refers to morphospecies found in studied Pineland ponds, while other data are from Levkov et al. (2016).

Species	**Width** (μm)	**Length** (μm)	**Striae** (in 10 μm)	**Ecology**
"*G. graciledictum/ subnaviculoides* NJ"	4.5-6	14-42	16-20	Circumneutral ponds
G. graciledictum	5.5-8	23-62	13-18	Meso-eutrophic lake with medium conductivity
G. hebridense	5.5-7	33-49	15-17	Oligotrophic slightly acidic waterbodies
G. naviculoides	7.5-10	26-69	11-13	Hypertrophic lake
G. subnaviculoides	4.5-6.5	16-39	16-22	Peat bogs

Several samples (from Slab Pond, NJPC46, 54, 55, 57, 60) had very high proportions of *G. graciledictum/ subnaviculoides* in valve counts. The *rbc*L data in most of these samples (Country Basin, Cardinal Basin, Leah Pond, Fig. 14) were dominated by the ASV 5+76+231 ("*G. affine/*

gracile /hebridense/ carolinense) cluster, four other ponds (Fletcher Basin, McDonald XPond, MacKay XPond, Muirfield Basin, Fig, 15) had a high percentage of ASV 8 (matched to *G. parvulum*), while samples from Slab Pond (Fig, 16) had an almost equal mixture of two clusters. In Dr. Enache's counts *Gomphonema* was split more finely, and the specimens in sample MacKay XPond were identified as *G. exilissimum* and in Muirfield Basin mostly as *Gomphonema* sp. 4 with fewer *G. parvulum*. As the Muirfield Basin sample had only ASV 8, either the specimens from the ASV 5+76+231 cluster were missed in this sample by the metabarcoding, or, more likely, the morphological variability of these taxa makes their separation with microscopy highly uncertain. At this point, it may only be concluded that the ASV 5+76+231 likely represent morphotaxa from the "*G. graciledictum /subnaviculoides"* species complex but establishing exact correspondence between morphology and barcodes requires culturing and Sanger sequencing.

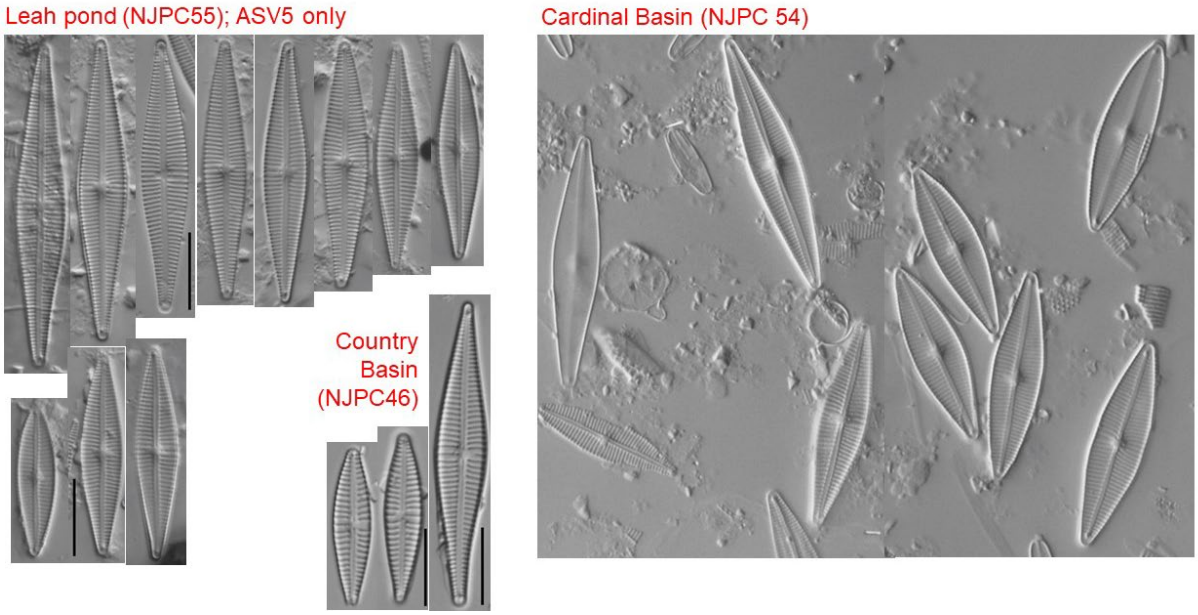

Figure 14. Specimens of the genus *Gomphonema* in samples dominated by ASV5+76+231 in *rbc*L metabacoding dataset.

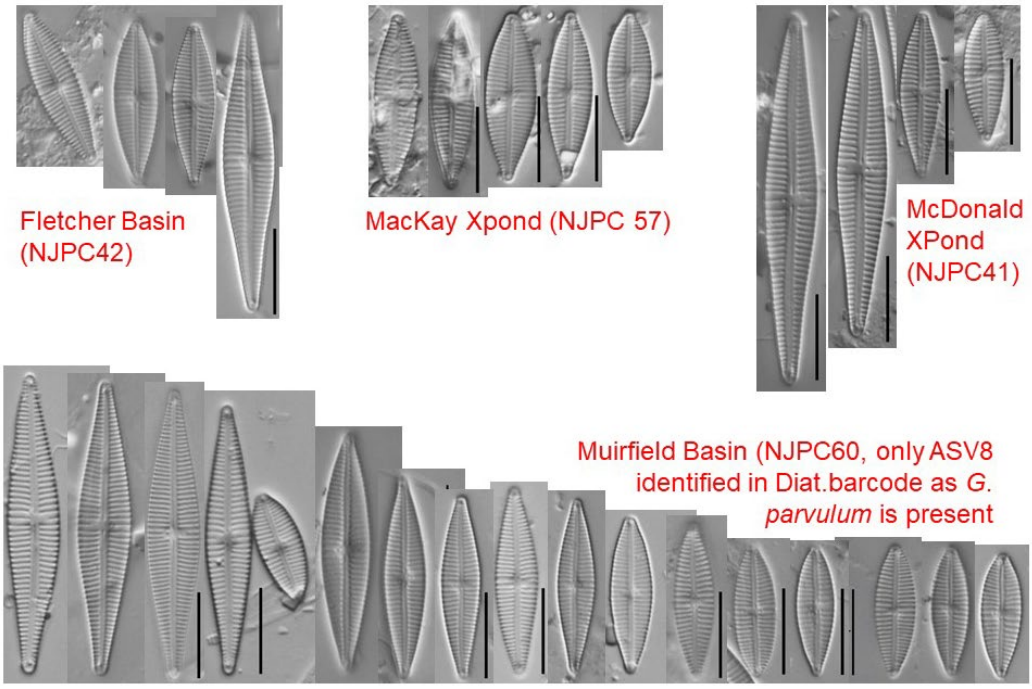

Figure 15. Specimens of the genus *Gomphonema* in samples dominated by ASV8 in *rbc*L metabacoding dataset.

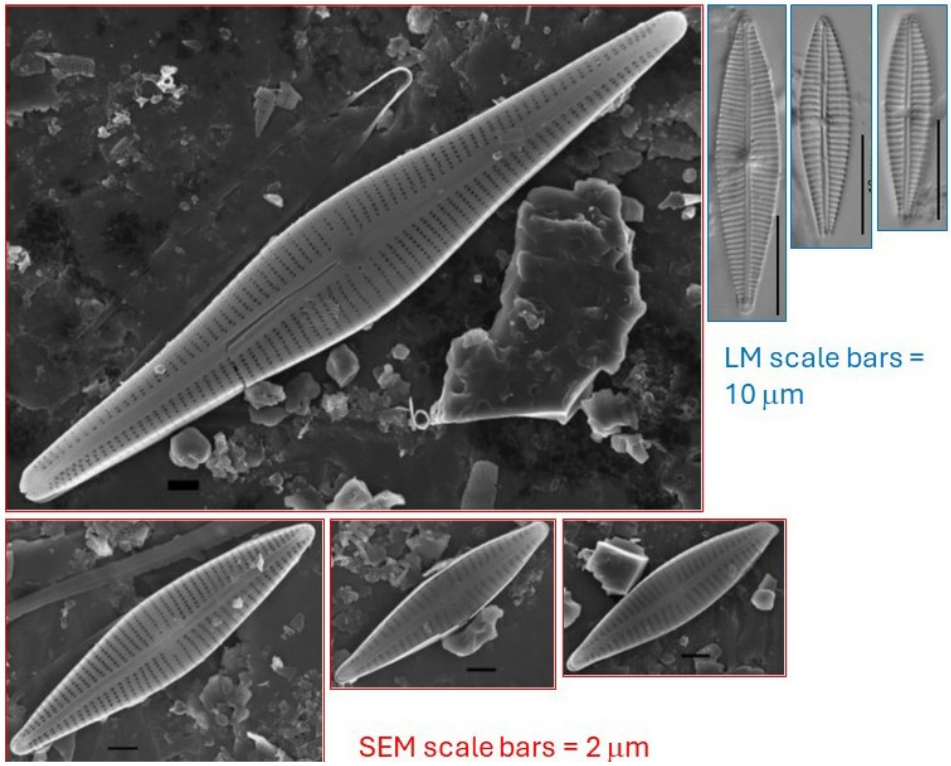

Figure 16. Specimen of *Gomphonema* from Slab Pond where approximately half of the rbcL reads belonged to the ASV5 (*G. graciledictum/ subnaviculoides* complex) and another half to ASV 8 (*G. parvulum*).

3.5.5. The genus Brachysira

Four species of *Brachysira* were detected in diatom counts with one species, *B. microcephala,* reaching high relative abundance in all samples from Goober Pond, lower numbers in the Mosquitofish Pond and a single valve recorded in Third XPond (NJPC48). Morphologically, the specimens from Goober and Mosquitofish ponds were indistinguishable (Fig. 17), but their barcode *rbc*L sequences differed by four basepairs. Both ASV 210 from Goober pond and ASV 1831from the Mosquitofish pond only differed by two basepairs from the sequence KU951597.1 accessioned as *Brachysira* sp. in the GenBank from an organism isolated in Vietnam and two sequences, one identified as *B. microcephala*, and another as *B. neoexilis* (which are synonyms) in the Diat.barcode reference database. The specimens from New Jersey ponds found in this project likely belong to yet undescribed species previously reported by Siver & Hamilton (2011) as Morphotype I of *Brachysira microcephala* common in New Jersey Pine Barrens and other locales in the eastern US with relatively acidic water bodies.

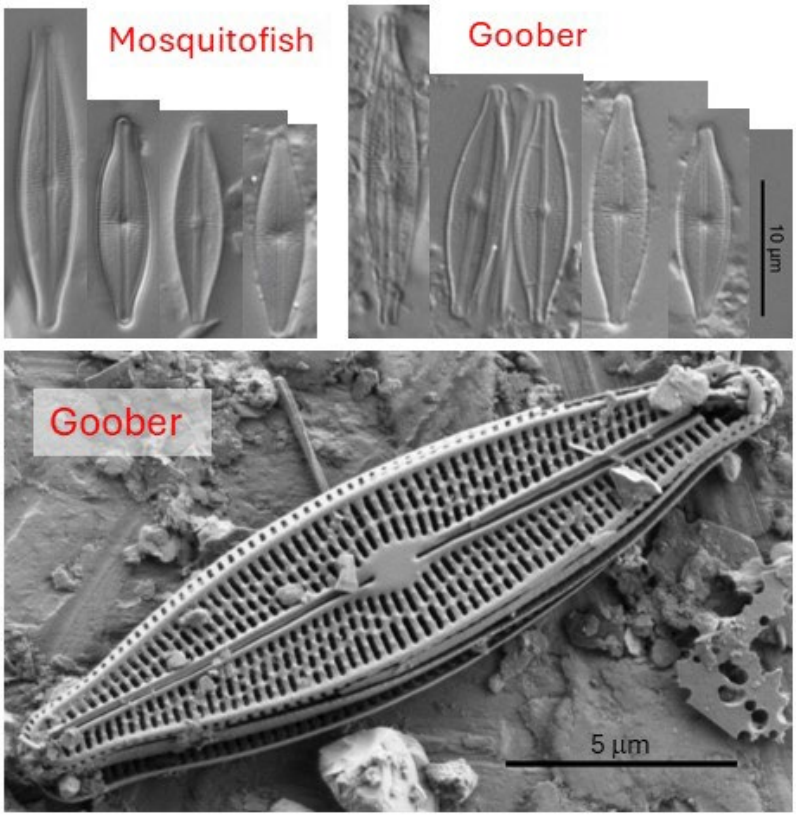

Figure 16. *Brachysira microcephala* morphotype I from Mosquitofish and Goober ponds. LM and SEM images.

3.5.6. Other genera

ASVs assigned to other genera were impossible to match unequivocally to morphotaxa, either because of the overall high sample diversity or species diversity within a particular genus. For example, samples with relatively high abundance of *Pinnularia, Stauroneis* or *Neidium* usually had several species of these genera. Several samples from studied ponds had high relative abundance of *Nitzschia* identified morphologically as *N. acicularioides* (Appendix 4), but several sequences apparently associatedwere mostly assigned to *N. palea* (the most abundant of these being ASV 11). In fact, 12 ASVs were assigned to *N. palea* while they were not very close genetically. This underscores the paucity of morphological characters separating species of *Nitzschia* that may be distantly related and indicate contrasting environmental conditions (Mann et al. 2021). Another species complex that proved too difficult to disentangle is *Frustulia saxonica/crassinervia*. There were several ASVs assigned to this species complex by Diat.barcode with ASV3 being the dominant in all studied samples where Frustulia speciemns were recorded based on microscopy. The other ASVs identified as *Frustulia saxonica/crassinervia* had minor contribution and therefore, it was impossible to assigned them to specific morphological entities.

4. Summary and Conclusions

DNA metabarcoding approach using two DNA markers, pan-eukaryotic 18S_V9 and diatom-specific *rbc*L, revealed highly diverse microbial eukaryotic assemblages in studied ponds. Numerous taxonomic groups of protists were characterized by 18S_V9 marker, while the overall diversity of diatom *rbc*L sequences was about a third higher than the diversity of diatom morphospecies. There could be various reasons for encountering multiple barcode sequences per morphological entity, such as cryptic diversity, intraspecific or intragenomic variability, although *rbc*L gene has not been known for significant intragenomic variation (Kelly et al. 2018). Metabarcoding is known for its high sensitivity (Keck et al. 2017) and the ability to detect even rare organisms compared to traditional microscopy approaches.

While water chemistry and biological sampling were separated in time, the differences in assemblage composition related to water quality were still evident. There was a clear gradient of assemblage composition along the pH gradient, and the influence of nutrient enrichment was also evident from the exploratory numerical analyses (Figs 3 and 5).

Environmental DNA metabarcoding has proven to be a cost-effective approach to biodiversity and water-quality assessment (Pawlowski et al. 2018) and this study confirms that it may be used successfully to monitor the health of various water bodies, including ephemeral ponds in New Jersey. As the cost of sequencing continuously declines, the taxonomic reference databases fill up, and laboratory procedures get standardized, this approach only gets more attractive with time. Importantly, the correspondence between ASVs and specific environmental conditions can be established without knowing the identity of organisms from which these sequences originated. This taxonomy-blind approach can be used for water-quality monitoring before the reference databases are complete (Rimet et al. 2018). Using wide-coverage markers, such as 18S_V9, in addition provides coverage of multiple groups of organisms which may respond to different stressors, which creates a wholistic picture of the microbial community.

Table 5. List of newly established diatom barcodes and their contribution in total reads number in the 61 sample *rbc*L dataset.

Morphospecies	**ASV #**	**% reads**
Discostella woltereckii	518	0.047
Ulnaria grunowii	42	1.076
Fragilaria cf. *capucina*	104	0.490
Tabellaria quadriseptata	188	0.233
Eunotia naegeli/genuflexa	1, 2, 68, 209, 254, 364, 435, 654, 1101	14.100
Eunotia pirla	43, 97	1.600
Eunotia metamonodon	15, 202, 454	2.280
Eunotia tauntoniensis	6, 138	4.170
Eunotia nymanniana	185	0.236
Gomphonema graciledictum/ subnaviculoides	5, 76, 231	5.330
Brachysira microcephala morphotype I	210, 1831	0.190

One goal of this project was to infer diatom barcodes from the environmental samples using relatively low-diversity assemblages from artificial substrates. Barcoding sequences were established with high certainty for 11 morphological entities (Table 5), but the overall sample diversity was not sufficiently low to establish a clear correspondence between the remaining majority of morphospecies and barcodes. In some cases, multiple highly similar sequences corresponded to a single morphological entity, while more often, there were multiple species from the same genus and multiple similar ASVs from the same genus found in the same samples. Nevertheless, the total contribution of reads that can be related to morphotaxa based on the new taxonomic assignments obtained in this study is 29.8% for the *rbc*L 61-sample dataset, which is

higher than 21% of reads that could have been assigned to species level with Diat.barcode. In addition, one of the newly established barcodes, corrects the assignment of *Ulnaria ulna* to *U. grunowii*. By combining the Diat.barcode and new assignments, 50% of reads in the studied samples can be taxonomically assigned to species level.

We conclude that metabarcoding is a valuable tool for biological monitoring of New Jersey waterbodies even in the absence of complete reference databases. Further investigations would improve the quality of taxonomic assignments either by establishing barcodes from environmental samples as was done in this project, or by standard culturing and Sanger sequencing approaches. Alternatively, sequences themselves can serve as environmental indicators in a taxonomy-blind approach. Either taxon-specific or wide-coverage markers can be easily used depending on the conservation targets and monitoring goals.

5. References

1. Alexson, E.E., Reavie, E.D., Van de Vijver, B., Wetzel, C., Ector, L., Wellard Kelly, H.A., Aliff, M.N. and Estepp, L.R. Revision of the needle-shaped Fragilaria species (Fragilariaceae, Bacillariophyta) in the Laurentian Great Lakes (United States of America, Canada) Journal of Great Lakes Research, **2022**, *48*, 999-1020.
2. Barta, B., Szabó, A., Szabó, B., Ptacnik, R., Vad, C.F. and Horváth, Z. How pondscapes function: connectivity matters for biodiversity even across small spatial scales in aquatic metacommunities. Ecography, **2024**, e06960. https://doi.org/10.1111/ecog.06960
3. Bruder,K. and Medlin,L.K. Molecular assessment of phylogenetic relationships in selected species/genera in the Naviculoid diatoms (Bacillariophyta). I. The genus *Placoneis*. Nova Hedwigia, **2007**, *85*, 331-352.
4. Bunnell, J.F., K.J. Laidig, P.M. Burritt, and Sobel M.C. Vulnerability and comparability of natural and created wetlands. Final report to the U.S. Environmental Protection Agency, **2018**. Pinelands Commission, New Lisbon, New Jersey, USA.
5. Gohl, D., Vangay, P., Garbe, J., MacLean, A., Hauge, A.; Becker, A., Gould, T.J., Clayton, J.B., Johnson, T.J.; Hunter, R., Knights, D., and Beckman, K.B. Systematic improvement of amplicon marker gene methods for increased accuracy in microbiome studies. *Nat. Biotechnol.* **2016**, *34*, 942–949. https://doi.org/10.1038/nbt.3601
6. Guillou, L., Bachar, D.; Audic, S., Bass, D., Berney, C.; Bittner, L., Boutte, C., Burgaud, G., de Vargas, C., Decelle, J., del Campo, J., Dolan, J. R., Dunthorn, M., Edvardsen, B., Holzmann, M., Kooistra, W. H. C. F., Lara, E., Le Bescot, N., Logares, R., and Christen, R. The Protist ribosomal reference database (PR^2): A catalog of unicellular eukaryote small sub-unit rRNA sequences with curated taxonomy. Nucleic Acids Res. **2013**, *41(Database issue),* D597–D604. https://doi.org/10.1093/nar/gks1160
7. Hill, M. J., Greaves, H. M., Sayer, C. D., Hassall, C., Milin, M., Milner, V. S., Marazzi, L., Hall, R., Harper, L. R., Thornhill, I., Walton, R., Biggs, J., Ewald, N., Law, A., Willby, N., White, J. C., Briers, R. A., Mathers, K. L., Jeffries, M. J., and Wood P. J. Pond ecology and conservation: research priorities and knowledge gaps. Ecosphere, **2021**, *12*, e03853. 10.1002/ecs2.3853
8. Keck, F., Vasselon, V., Tapolczasi, K., Rimet, F., and Bouchez, A. Freshwater biomonitoring in the information age. Front. Ecol. Environment, **2017**, *15*, 266–274. https://doi.org/10.1002/fee.1490

9. Kelly, M., Boonham, N., Juggins, S., Killie, P., Mann, D., Pass, D., Sapp, M., Sato, S. and Glover, R. A. DNA based diatom metabarcoding approach for water framework directive classification of rivers. Report No. SC140024/R, Environment Agency, **2018**, 146 pp. https://www.gov.uk/government/publications/a-dna-based-metabarcoding-approach-to-assess-diatom-communities-in-rivers
10. Kermarrec, L.; Franc, A.; Rimet, F.; Chaumeil, P.; Humbert, J.F.; Bouchez, A. Next-generation sequencing to inventory taxonomic diversity in eukaryotic communities: a test for freshwater diatoms. Mol. Ecol. Resour., **2013**, *13*, 607–619. https://doi.org/10.1111/1755-0998.12105
11. Kochoska, H., Chardon, C., Chonova, T., Keck, F., Kermarrec, L., Larras, F., Lefrancois, E., Rivera, S.F., Tapolczai, K., Vasselon, V. and Levkov, Z. Filling reference libraries with diatom environmental sequences: strengths and weaknesses. Diatom Research, **2023**, *38*, pp.103-127.
12. Krammer, K., and Lange-Bertalot, H. Bacillariophyceae. 2. Teil: Bacillariaceae, Epithemiaceae, Surirellaceae. In: Susswasserflora von Mitteleuropa; Ettl, H., Gerloff, J., Heynig, H., Mollenhauer, D., Eds.; Gustav Fisher Verlag: Jena*, Germany,* **1986-2004**; *Bands 2/2-4.*
13. Kulikovskiy, M., Lange-Bertalot, H., Annenkova, N., Gusev, E., and Kociolek, J.P. Morphological and molecular evidence support description of two new diatom species from the genus *Ulnaria* in Lake Baikal. Fottea, **2016**, *16*, 34–42.
14. Lange-Bertalot, H., Bak, M., Witkowski, A., and Tagliaventi, N. *Eunotia* and some related genera. Diatoms of the European Inland Waters and Comparable Habitats, **2011,** *6, 1-747.*
15. Lange-Bertalot, H., and Ulrich, S. Contributions to the taxonomy of needle–shaped Fragilaria and *Ulnaria* species. Lauterbornia, **2014**, *78*, 1–73.
16. Levkov, Z., Mitić-Kopanja, D., and Reichardt, E. The diatom genus *Gomphonema* from the Republic of Macedonia. Diatoms of Europe. Diatoms of the European Inland Waters and Comparable Habitats*,* **2016**, *8, 1-552.*
17. Mann, D.G., Trobajo, R., Sato, S., Li, C.L., Witkowski, A., Rimet, F., Ashworth, M.P., Hollands, R.M. and Theriot, E.C. Ripe for reassessment: A synthesis of available molecular data for the speciose diatom family Bacillariaceae. Molecular Phylogenetics and Evolution*,* **2021**, *158(106985), 1-19.*
18. Martin, M. Cutadapt removes adapter sequences from high-throughput sequencing reads. EMBnet.journal, **2011**, *17*, 10–12. https://doi.org/10.14806/ej.17.1.200

19. Morien, E., Parfrey, L.W. SILVA v128 and v132 dada2 formatted 18s 'train sets' (1.0) [Data set]. Zenodo. 2018. Available online: https://doi.org/10.5281/zenodo.1447330 (accessed 1.10.2023).
20. Patrick, R.M., and Reimer, C.W. The Diatoms of the United States exclusive of Alaska and Hawaii, V. 1. Monographs of the Academy of Natural Sciences of Philadelphia, **1966**, *13.*
21. Pérez-Burillo, J., Valoti, G., Witkowski, A., P. Prado, P., Mann, D.G., and. Trobajo, R. Assessment of marine benthic diatom communities: insights from a combined morphological–metabarcoding approach in Mediterranean shallow coastal waters. Mar. Pollut. Bull., **2022**, *174*, 113183. https://doi.org/10.1016/j.marpolbul.2021.113183
22. Pawlowski. J., and Kelly-Quinn, M., Altermatt, F. *et al.* The future of biotic indices in the ecogenomic era: Integrating (e) DNA metabarcoding in biological assessment of aquatic ecosystems. Sci Total Environ., **2018**, *637-638*,1295-1310. https://doi.org/10.1016/j.scitotenv.2018.05.002
23. Ponader, K.C., Charles, D.F., Belton, T.J. *et al.* Total phosphorus inference models and indices for coastal plain streams based on benthic diatom assemblages from artificial substrates. Hydrobiologia, **2008**, *610*, 139–152. https://doi.org/10.1007/s10750-008-9429-6
24. R Core Team. R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria, **2023**. Available online: https://www.R-project.org/ (accessed 12.12.2023).
25. Rimet, F., Abarca, N., Bouchez, A., Kusber, W.H., Jahn, R., Kahlert, M., Keck, F., Kelly, M.G., Mann, D.G., Piuz, A. and Trobajo, R. The potential of High-Throughput Sequencing (HTS) of natural samples as a source of primary taxonomic information for reference libraries of diatom barcodes. *Fottea* **2018**, 18, 37-54.
26. Rimet, F., Chonova, T., Gassiole, G., Kahlert, M., Keck, F., Kelly, M., Kulikovskiy, M., Mann, D., Pfannkuchen, M.A., Baričević, A., Trobajo, R., Vasselon, V., Zimmermann, J., Wetzel, C.E., and Bouchez, A. Diat.barcode: a DNA tool to decipher diatom communities for the evaluation environmental pressures. ARPHA Conference Abstracts **2021**, *4*, e64940. https://doi.org/10.3897/aca.4.e64940
27. Rimet, F., Gusev, E., Kahlert, M., Kelly, M. G., Kulikovskiy, M., Maltsev, Y., Mann, D. G., Pfannkuchen, M., Trobajo, R., Vasselon, V., Zimmermann, J., and Bouchez, A. Diat.Barcode, an open-access curated barcode library for diatoms. *Sci. Rep.-UK* **2019**, *9*, 1. https://doi.org/10.1038/s41598-019-51500-6

28. Saitou, N., and Nei, M. The neighbor-joining method: A new method for reconstructing phylogenetic trees. *Molecular Biology and Evolution,* **1987**, *4*, 406-425.
29. Schultz, K., Dressler, M., Jaques, O., Springer, A.., Frank, M., and Hübener, T. Hidden complexity: An assessment of species diversity within the genus Discostella (Bacillariophyta). *Fottea, Olomouc,* **2024**, *24*, 42-60.
30. Siver, P.A., and Hamilton, P.B. Diatoms of North America: The Freshwater Flora of Waterbodies on the Atlantic Coastal Plain. *Iconographia Diatomologica,* 2011, *22,* 1-916.
31. Spaulding, S., Potapova, M.G., Bishop, I.W., Lee, S.S., Gasperak, T.S., Jovanoska, E., Furey, P.C., and Edlund, M.B. *Diatoms.org*: supporting taxonomists, connecting communities. *Diatom Res*. **2021**, *36*, 291–304. https://doi.org/10.1080/0269249X.2021.2006790
32. Stoeck, T., Bass, D., Nebel, M., Christen, R., Jones, M. D. M., Breiner, H.-W., and Richards, T. A. Multiple marker parallel tag environmental DNA sequencing reveals a highly complex eukaryotic community in marine anoxic water. *Mol. Ecol.* **2010**, *19 Suppl 1*, 21–31. http://doi.org/10.1111/j.1365-294X.2009.04480.x
33. Tamura, K., Stecher G., and Kumar S. MEGA 11: Molecular Evolutionary Genetics Analysis Version 11. Molecular Biology and Evolution, 2021, https://doi.org/10.1093/molbev/msab120.
34. Turk Dermastia, T.; Vascotto, I.; Francé, J., Stanković, D., and Mozetič, P. Evaluation of the *rbcL* marker for metabarcoding of marine diatoms and inference of population structure of selected genera. *Front. Microbiol*. **2023**, *14*, 1071379. https://doi.org/10.3389/fmicb.2023.1071379
35. Vasselon, V., Rimet, F., Tapolczai, K., and Bouchez, A. Assessing ecological status with diatoms DNA metabarcoding: Scaling-up on a WFD monitoring network (Mayotte island, France). *Ecol. Indic*. **2017**, *82*, 1–12. https://doi.org/10.1016/j.ecolind.2017.06.024
36. Walker, W.C., and Chase H.H. *Notes on some new and rare Diatoms. Photoplates. Series I.* **1886**, pp. 1-6, 2 pl. Utica, New York: Curtiss & Chlids Print
37. Zampella, R.A., Laidig K.J., and Lowe, R.L. Distribution of diatoms in relation to land use and pH in blackwater coastal plain streams. Environ Management, **2007**, *39*, 369-84. https://doi.org/10.1007/s00267-006-0041-0