A two-decade microbial time series from a freshwater lake, introducing the limony and TYMEFLIES datasets

Authors

Robin R. Rohwer[1], Katherine D. McMahon[2,3]

Affiliations
1. Department of Integrative Biology, The University of Texas at Austin, Austin, Texas, USA
2. Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin, USA
3. Department of Civil and Environmental Engineering, University of Wisconsin-Madison, Madison, Wisconsin, USA
correspondence: Robin Rohwer (robin.rohwer@gmail.com)

Abstract

How microbial communities change over time is key to understanding how ecosystems will respond to global change, and to answering fundamental questions about how microbial ecology and evolution unfold. However, our understanding of microbial change is limited by a lack of long-term observations. Using archived filters from freshwater Lake Mendota (WI, USA), we created a 20-year microbial time series that begins in year 2000, a decade before Illumina DNA sequencing hit the market. We characterized 1,023 samples and controls with 16S rRNA gene amplicon sequencing in the "limony" dataset, and 471 samples and controls with shotgun metagenome sequencing in the "TYMEFLIES" dataset. In addition to the raw sequencing data, we point users to associated data products: paired environmental data collected by the North Temperate Lakes Long-Term Ecological Research program (NTL-LTER), our R package to work with curated limony data, and metagenomic assemblies and metagenome-assembled genomes (MAGs) created from the TYMEFLIES dataset. The limony and TYMEFLIES datasets, along with their paired environmental data and reference genomes, are a unique and ready-to-use resource for diverse scientists including limnologists, ecologists, evolutionary biologists, bioinformaticians, modellers, and educators.

Background & Summary

Often considered the birthplace of limnology, Lake Mendota (WI, USA) is one of the best-studied lakes in the world. Limnological records date back to the early 1900s, and in recent decades this research has expanded to include lake microbes. Here we describe in detail the collection and sequencing of two 20-year microbial time series: a 1,023-sample 16S rRNA gene amplicon dataset we refer to as limony and a 471-sample metagenome dataset we call TYMEFLIES. We also provide background on the lake, describe an R package for working with processed limony data, and point readers to complimentary data products like environmental data and metagenome-assembled genomes (MAGs). We hope this background and practical guidance will enable other scientists to analyse limony and TYMEFLIES in creative ways that are conscious of the data's limitations while taking advantage of its strengths.

Lake Mendota

Lake Mendota is in southern Wisconsin, on the shores of the University of Wisconsin-Madison and the Wisconsin state capitol (Fig 1). Lake Mendota is dimictic and temperate, meaning that it has a thermally stratified water column that mixes vertically twice a year in spring and fall, and the lake freezes in the winter. In a typical year, the ice-on season lasts from late December to early April, with ice thickness reaching 40 cm (Magee et al. 2016). Vertical mixing occurs immediately after ice-off in April and as temperatures cool in October, and hypolimnetic anoxia develops during the summer stratified period (Ladwig et al. 2021). Summer surface

water temperatures typically reach 25°C, and the lake is popular for recreational activities such as fishing, sailing, and swimming. In early spring (mid-April to mid-June) a clearwater phase with typical water clarities of 7-10 m delineates a shift in phytoplankton community from diatoms to Cyanobacteria (Carey et al. 2016; Matsuzaki et al. 2020). However, Lake Mendota is eutrophic meaning that it has high concentrations of nitrogen and phosphorus, and summer water clarities of 1-2 m are punctuated by surface cyanobacterial blooms, which prevent recreation due to cyanotoxins and noxious odours (Lathrop et al. 1996; Lathrop 2007; Lathrop and Carpenter 2014; Beversdorf et al. 2015, 2017). Lake Mendota is considered a fairly typical lake, although at 39.6 km^2 it is on the larger side, and the maximum fetch of 6 miles results in significant wave action. Lake Mendota is not unusually deep, with a maximum depth of 25 m and an average depth of 12.7 m (Brock 1985).

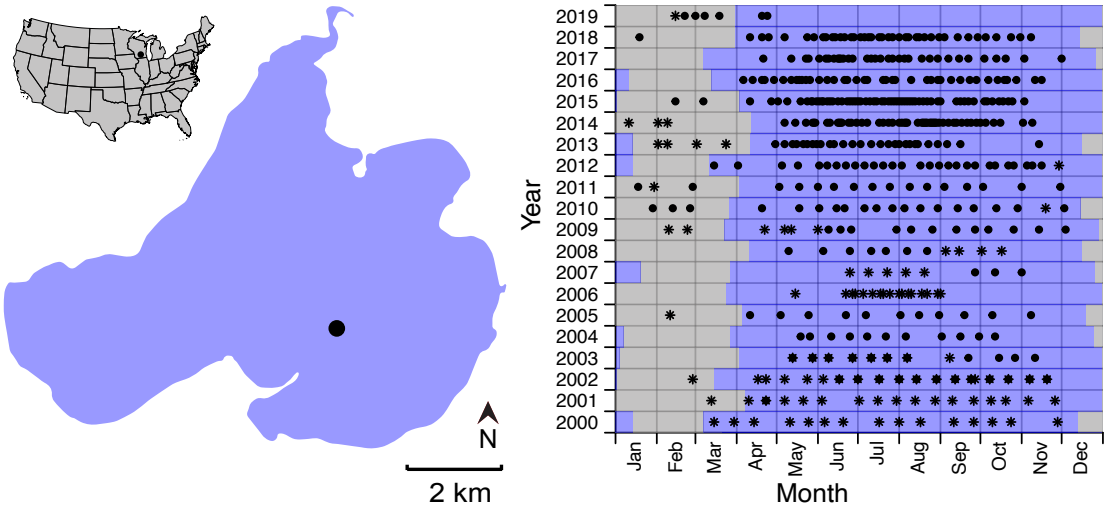

Fig 1. Sample dates and location. A: Map of Lake Mendota, with the deep hole sampling location indicated with a dot. B: Sample dates are displayed as points. White shading indicates periods of ice-on and blue shading indicates periods of ice-off. Samples points indicated with a star as opposed to a solid dot were collected with non-standard filtration, depth, or location (Details in Online-only Table 1). Note that some dates include both standard and non-standard samples.

Long-term research on Lake Mendota (WI, USA) began in the early 1900s (Brock 1985; Lathrop et al. 1996) and today includes research programs led by the University of Wisconsin-Madison's Center for Limnology (CFL) and Environmental Chemistry and Technology program (EC&T), the Wisconsin Department of Natural Resources (WI DNR), the United States Geological Survey (USGS), and the North Temperate Lakes Long Term Ecological Research program (NTL-LTER). A microbial observatory (NTL-MO) (Kane 2004) was established on Lake Mendota in 1999 with funding through the "Microbial Observatories" program at the National Science Foundation, and the NTL-LTER now supports continued routine collection of microbial samples that can be used for a variety of analyses including Illumina-based 16S rRNA gene amplicon sequencing to yield "iTAGs" and shotgun sequencing to yield metagenomes. As a result of this intensive study, a wealth of long-term environmental data is publicly available through public sources such as the Environmental Data Initiative (EDI) and the USGS Water Data for the Nation.

The lake iTag measurements over nineteen years dataset, known by the acronym limony, and the Twenty Years of Metagenomes Exploring Interannual Eco/evo Shifts dataset, known by the acronym TYMEFLIES, span the duration at the time of sequencing of the archived Lake Mendota Microbial Observatory, from spring 2000 to spring 2019 (Fig 2). During this time span

Lake Mendota has experienced the long-term press of climate change, observable as shorter ice cover duration, longer stratification periods, increased phosphorus loading from heavy rains, and shifting phenology (Carpenter et al. 2015; Magee et al. 2016, 2019; Matsuzaki et al. 2020; Ladwig et al. 2022). However, during the two decades covered by the limony dataset Lake Mendota has also experienced abrupt ecological shifts due to aquatic invasive species (Vander Zanden et al. 2024). In 2009, a population explosion of the invasive invertebrate predator spiny water flea (*Bythotrephes longimanus*) resulted in nearly a meter loss of water clarity due to a trophic cascade in which zooplankton predation reduced grazing pressure on phytoplankton (Walsh et al. 2016a; b, 2017, 2019), as well as shifting penology of the spring clearwater phase (Matsuzaki et al. 2020) an increase in hypolimnetic anoxia (Rohwer et al. 2024b). In the fall of 2015, invasive zebra mussels (*Dreissena polymorpha*) were first detected, and by the summer of 2016 these invasive filter feeders were already covering most hard surfaces in the near-shore zone, resulting in a 300% increase in zoobenthos and phytobenthos abundance by 2018 (Spear et al. 2022). However, zebra mussels have not resulted in clearer water despite theoretically filtering the epilimnion volume every 45 days (Spear et al. 2022), and they have resulted in increased cyanotoxins (Rohwer et al. 2023).

The microbial community of Lake Mendota has been extensively studied prior to the limony and TYMEFLIES datasets, so broad trends about Lake Mendota's microbial community are well established. Early studies of community composition used ARISA fingerprinting, which differentiates organisms based on the variable length of the 16S-23S rRNA gene intergenic spacer region (Fisher and Triplett 1999). These ARISA-based studies found consistent seasonal changes in bacterial community composition (Yannarell et al. 2003; Shade et al. 2007; Kara et al. 2013). Analyses of early Illumina 16S rRNA gene amplicon sequencing confirmed the strong seasonality of many Lake Mendota microbes, and identified biotic interactions as likely drivers of community change (Hall et al. 2017; Herren and McMahon 2017). In limony and TYMEFLIES, microbial phenology remains a major dynamic, with seasonal patterns observed in bacteria (Rohwer et al. 2023), including at the strain level (Rohwer et al. 2024a), as well as among protists (Krinos et al. 2024) and viruses (Zhou et al. 2024). Overlaying these short term patterns, the limony and TYMEFLIES datasets have identified long-term shifts concurrent with species invasions (Rohwer et al. 2023; Krinos et al. 2024) and environmental extremes (Rohwer et al. 2024a). Given the complexity of these datasets and the breadth of available environmental data, many more microbial trends and drivers likely yet to be discovered.

limony: 980 iTags from 477 Sample Dates

The limony dataset is unique because of the length of its time span and sampling frequency, and because most sample dates include iTags from two biological duplicates (Fig 1). During the first decade of the time series, samples were typically collected fortnightly throughout the ice-off season, and during the second decade samples were typically collected weekly. These frequent observations are valuable for documenting community succession and revealing possible interactions over time because bacterial communities turnover on short time scales. For example, in surface seawater generation times may vary from hours to several days (Bunse and Pinhassi 2017). Regardless of generation times, multi-decadal time series are needed to tease apart true long-term change from large underlying multi-scale variability (White 2019).

The existence of biological duplicates (separate water grabs) also makes limony a powerful and unique resource. Community-derived 16S rRNA gene amplicon data are often criticized for being noisy, as evidenced by the many methods that have been created to cluster, filter, and error correct it (Eren et al. 2015; Callahan et al. 2016, p. 2; Edgar 2016; Amir et al. 2017; Schloss 2020). However, environmental data are also often highly variable at multiple scales. The NTL-MO collects biological duplicates that we sequenced in parallel, so they capture both

the technical noise, which is often addressed in various data processing pipelines, and the environmental noise, which is typically ignored in microbial ecology (Prosser 2010; Lennon 2011).

TYMEFLIES: 471 metagenomes from 463 Sample Dates

The TYMEFLIES dataset is similarly unique for its length and sample density, and also because of the depth of metagenome sequencing. TYMEFLIES samples include a single replicate from all limony sample dates with enough DNA yield to perform shotgun metagenome sequencing. Several duplicated samples exist due to a repeated "generous donor" sample included on all 5 sequencing plates, and to a few instances where sampling was nonstandard and differed between replicates. TYMEFLIES metagenome sequencing depths are 80 $\pm$ 20 million reads/sample and 23 $\pm$ 6 billion bases/sample. Deep sequencing such as this improves recovery of metagenome-assembled genomes (MAGs), and it enables genomic analyses of recovered MAGs that require sufficient depth for single nucleotide variant (SNV) calling. SNV-calling, for example, allows for strain-resolved analyses that can be used to identify evolutionary processes.

Methods

Sample Collection

Collection and transport. Water samples were collected from a central point in Lake Mendota (43°05'58.2"N 89°24'16.2"W), at the same location as a weather buoy that has collected basic physiochemical measurements during ice-off season since 2006 (Fig 1). This location is the deepest part of the lake (25 m) and is over 1.4 km from shore in all directions. Integrated 12 m samples (approximating epilimnion depth) were collected by lowering a weighted 12 m tube, plugging the top of the tube to pull out a vertical section of the water column, and then mixing this depth-integrated water sample in a 2 L bottle (Fig 2). Duplicate samples were collected from opposite sides of the boat. In the winter, samples were collected from a hole drilled by an ice augur. Some limony samples deviate from this exact method, such as winter samples that were surface water grabs from closer to shore, or samples in 2006-2007 that were depth integrated based on the measured epilimnion. These non-standard samples are indicated by stars in Fig 1 and detailed in Online-only Table 1.

Filtering and storage. Water was transported back to the laboratory for filtering, which could take between 15 min to an hour, depending on how quickly the field work was performed and whether the samples were transported across campus by bike or by walking. Water samples were filtered onto 0.22 μm polyethersulfone filters (Pall Corporation), and the filters were placed in airtight screw-cap tubes and stored in a -80°C freezer until DNA extraction during 2018-2019 (Fig 2). During years 2000-2003, one hundred and nine samples were collected onto 0.22 μm filters after passing through a 10 μm prefilter. These non-standard samples are indicated by stars in Fig 1 and detailed in Online-only Table 1.

DNA Extraction

Duplicates. Two filters were sequenced per sample date, one from each biological replicate (duplicate water grab). Eighty-eight sample dates deviate from this scheme however. For forty-six of these samples, two filters collected from the same water grab were used as duplicates. The remaining forty-two samples lack replicates altogether because additional filters were not available. These non-standard samples are detailed in Online-only Table 1.

Randomization. Filters were put in random order before performing DNA extractions and sequencing (Fig 2). Duplicates were split apart into two groups, each group's order was randomized separately, and extractions were performed in the order of group 1 followed by

group 2. For iTag analysis, each group was spread over two sequencing plates, so that duplicates were sequenced in separate Illumina runs. For metagenome analysis, the replicate with higher DNA yield was chosen and the random order was generally maintained. Samples were never ordered by date, but for example low-yield samples were grouped into the same sequencing run which was subject to manual pipetting at JGI. Extraction and sequencing orders are detailed in Online-only Table 1.

Extractions. DNA extractions were performed by hand using FastDNA spin kits with Lysing Matrix A used for bead-beating (MP Biomedicals) (Fig 2). To improve cell lysis during bead beating, folded filters were cut into strips using flame-sterilized scissors and tweezers, and only half of each filter was included in bead beating and DNA extraction. In preliminary testing, we found we obtained the same or higher yield from half filters, which we attributed to less-full sample tubes improving the efficacy of bead beating. The first and last extraction of every 100-extraction kit was performed on a sterile filter as a blank. All extractions were performed by the same person (RRR) according to the same method, and no differences in DNA yield were observed by extraction date or kit. Extraction yields are provided in Online-only Table 1.

Creating the limony dataset

16S rRNA gene amplicon sequencing. Sequencing was performed at Argonne National Lab according the procedure established by Walters *et al.* (2016) and recommended by the Earth Microbiome Project. The V4-V5 variable region was amplified (515F-Y GTGYCAGCMGCCGCGGTAA, 926R CCGYCAATTYMTTTRAGTTT), and sequencing was performed on an Illumina MiSeq instrument using version 2 chemistry. Sequencing was split across 4 plates that multiplexed 245, 247, 269, and 262 samples each. The sequencing plate for each sample is provided in Online-only Table 1.

Quality filtering and demultiplexing. Before demultiplexing, index reads were quality filtered using vsearch (Rognes et al. 2016) (vsearch v2.13.3) to retain only index reads where the expected errors were ≤ 0.03 (`--fastq_maxee = 0.03`). Mothur was then used to remove the matching sequences for the forward and reverse read files (Schloss 2020) (mothur v.1.44.1) (`list.seqs()` and `get.seqs()`). Quality control of index reads can remove cross-contamination that occurs during sequencing, sometimes called cross-talk or index hopping. This is not typically considered a major concern with the MiSeq platform (Larsson et al. 2018; van der Valk et al. 2018), but it has been observed at low levels ranging from 0.16 - 0.2% (Nelson et al. 2014; D'Amore et al. 2016). We chose a maximum expected errors cut-off of 0.03, to approximate the mean quality score cutoff of 25 suggested by Wright and Vestigian (2016).

Read quality cut-offs were chosen for each of the four limony sequencing plates separately. Cut-offs for Read quality were chosen by running vsearch (Rognes et al. 2016) with a range of expected error cutoffs and truncation lengths (`--fastq_trunclen`), and choosing a cutoff of acceptable quality and length that retained the most sequences. A maximum expected errors of 2 (`--fastq_maxee = 2`) was applied to all forward and reverse reads following truncation of the forward reads at 247 bp (plates 1 and 2) or 250 bp (plates 3 and 4) and truncation of the reverse reads at 195 bp (plates 1 and 2) or 192 bp (plates 3 and 4). Sequences were demultiplexed using qiime (Estaki et al. 2020) (qiime2-2019.4) without golay correction of the barcode sequences.

Error correction using dada2. For dada2 error correction (dada2 version 1.12.1) (Callahan et al. 2016), learning errors and pooled error correction were run separately on each plate, as recommended by the dada2 big data workflow tutorial. After removing phiX and merging forward and reverse reads, the four plates were combined for chimera detection.

Taxonomy assignment using TaxAss with the FreshTrain and Silva. We assigned taxonomy with the Silva database (Quast et al. 2013) (silva138) and the FreshTrain database (Newton et al. 2011) (FreshTrain15Jun2020silva138) using the TaxAss workflow (Rohwer et al. 2018) (v2.0.0) with a 98 percent identity cut-off. In addition to classifying freshwater heterotrophs with the FreshTrain nomenclature, TaxAss also adjusts the names in the Silva database to make duplicate names unique, and it renames the unknown sequences with the uniform format of `unnamed.NearestParentName`. Unique names within and between levels are required by our limony R scripts. Additionally, *Cyanobacteria* nomenclature from silva138 was manually edited to remove strain names from the genus-level assignments. These strain names were instead indicated at the species level, a level that is not included in the Silva database. The edited *Cyanobacteria* nomenclature is available on the TaxAss GitHub repository (https://github.com/McMahonLab/TaxAss).

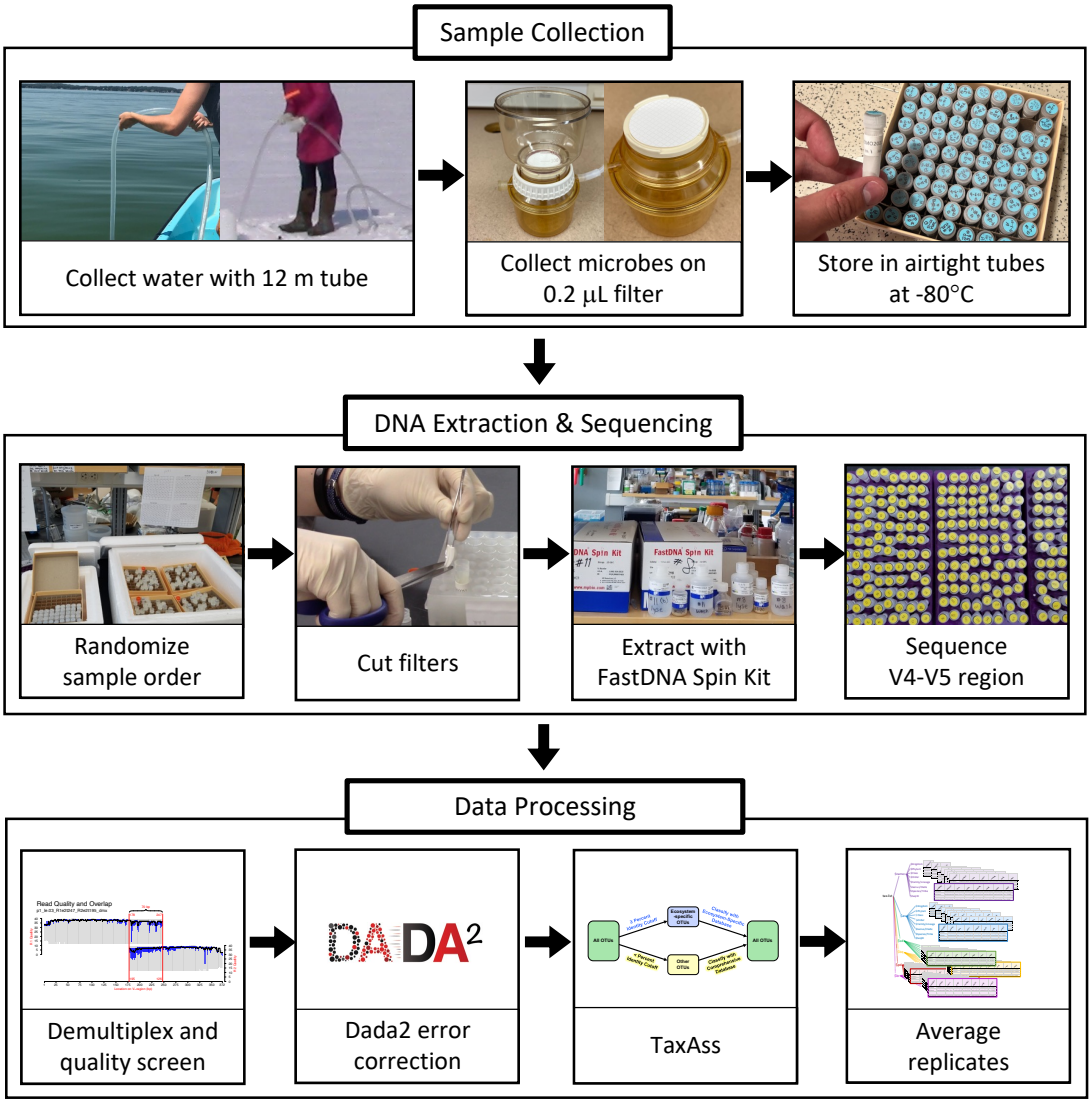

Fig 2. Methods overview. A schematic overview of the sampling methods, wet lab methods, and computational methods used to create the limony dataset.

Removing samples with low sequencing depth. The distribution of reads per sample approximated a normal distribution, and we chose a cut-off that maximized the number of samples kept, prioritizing the existence of duplicates over sequencing depth. A cut-off of 6000

reads per sample removed only 10 samples and maintained the resolution of a single read at 0.017% relative abundance. We did not rarefy the data, following the recommendation of McMurdie and Holmes (2014).

Processing in R. Custom R scripts were used to further process the samples by averaging replicates, calculating standard deviation, calculating several metrics of quantitation limit, and propagating these calculations to different taxonomic groupings. The exact annotated scripts are available at github.com/rrohwer/limony, and users can repeat this processing with different choices about library size or abundance cut-offs. After separating controls and applying cut-offs, the remaining replicate samples were combined. Mean abundances were calculated, along with standard deviations and information about whether an OTU was present in all replicates or at an abundance below the resolution of the smallest library size. After calculating these values at the OTU level, OTUs were grouped for each taxa level. Average abundances were summed, standard deviations were calculated by error propagation, and boolean information about replication and resolution was applied to all constituents at each taxon level. The result of this processing is stored as a list of each metric, with sub-lists for each taxon level (Fig 3). More details on the processed data format are available in the Usage Notes section.

Creating the TYMEFLIES dataset

Sample preparation. The DNA used for TYMEFLIES metagenome sequencing came from the same sample extractions as those used for the limony iTag dataset. However, in most cases only one biological replicate was sequenced with shotgun metagenomics. Since more DNA is required for metagenome sequencing, this replicate was typically chosen based on its higher DNA yield. Every sample date in the limony dataset with high enough DNA yield was also sent for metagenome sequencing. Online-only Table 1 can be used to link TYMEFLIES metagenomes with the matching limony iTag sample from the same DNA extraction.

Shotgun metagenome Sequencing. Samples were sent to the Department of Energy's Joint Genome Institute (JGI) for shotgun metagenome sequencing. A NovaSeq 6000 Sequencing System was used with an S4 flow cell to perform paired-end shotgun sequencing. The 471 samples were sequenced to a read depth of 80 $\pm$ 20 million reads/sample (23 $\pm$ 6 billion bases/sample). Samples were randomly ordered across across 5 plates submitted with 93, 93, 93, 93, and 74 samples each, as well as 30 low-DNA tubes requiring special manual sequencing preparation. Each of these 6 submissions included a repeat generous donor control sample. The sequencing plate for each sample is provided in Online-only Table 1.

Data Records

limony

Raw sequencing data for limony is available from the NCBI Sequence Read Archive under BioProject accession number PRJNA846788. A key linking sequencing codes to sample names and NCBI Run accessions (SRR#) is included as part of Online-only Table 1. The processed and curated limony dataset is available for download as an R package at github.com/rrohwer/limony.

TYMEFLIES

Raw sequencing data for TYMEFLIES was submitted to the NCBI Sequence Read Archive by the JGI Integrated Microbial Genomes & Microbiomes (IMG/M) pipeline, so each sample has a different BioProject accession. However, the Umbrella BioProject PRJNA1056043 links all TYMEFLIES samples as well as some associated assemblies and MAGs. A key with individual BioProject numbers, as well as sequencing codes, sample names and NCBI Run accessions (SRR#) is included as part of Online-only Table 1.

Environmental

The North Temperate Lake Long-Term Ecological Research program (NTL-LTER) collects routine limnological measurements, and their data products are available through the Environmental Data Initiative (EDI) (https://environmentaldatainitiative.org).

Some environmental data collected directly alongside the microbial samples is also available from the Environmental Data Initiative (EDI) (https://environmentaldatainitiative.org). Secchi disk water clarity measurements can be found under accession number knb-lter-ntl.416.1 and water temperature and dissolved oxygen measurements can be found under accession number knb-lter-ntl.415.2 (Rohwer and McMahon 2022a; b). Data packages from EDI are easy to access and can be downloaded directly into R, python, or matlab, or downloaded as a plain text file.

Technical Validation

Randomization, Controls, Mock Communities, and Duplicates

The order of sample extractions and sequencing was randomized, with limony duplicates placed on separate sequencing plates. The first and last extraction of every 100-sample prep kit was a clean filter. These 22 extraction blanks were spread across the 4 limony MiSeq plates, and can be used to detect kit contamination, cross contamination, and sequencing cross-talk. Each of the four limony sequencing plates also included 2 mock communities and 3 repeated "generous donor" samples to control for inconsistencies in community composition as observed by Yeh *et al.* (2018). The two mock communities were an even and a staggered marine mock community provided by the Fuhrman Lab (Yeh et al. 2018). A generous donor control is a repeated sample, which is less controlled than a mock community but is more representative of the actual community composition, diversity, and evenness. Three generous donors were created from different seasons to encompass a range of lake communities. These were collected with the same methods as the limony samples, but more filters were extracted and the extracted DNA was mixed before being divided into aliquots. The generous donor samples are from May 15, 2014, August 26, 2014, and November 8, 2018. The TYMEFLIES sequencing included the November 8, 2018 generous donor sample as part of all 6 individual sample submissions, and sequencing plates also included the standard blank wells used by JGI's internal quality control pipeline.

Usage Notes

Sample Metadata

A table that lists sample names, sample dates, sequencing codes (as identified on NCBI SRA), and whether samples were taken with non-standard depth, filtration, or location is available as Online-only Table 1.

Additional Microbial Resources

A manually curated taxonomy of abundant freshwater heterotrophs was created by Newton *et al.* using full length clone libraries (Newton et al. 2011). This Freshwater Training set (the FreshTrain) enables fine-resolution taxonomy classifications of iTag data when used with a Taxonomy Assignment tool (TaxAss) to classify 16S rRNA gene amplicon data (Rohwer et al. 2018). We encourage users of limony to take advantage of this unique taxonomy resource so that their results are easily comparable to other freshwater datasets, and so that they can perform taxonomic analyses at ecologically relevant levels. Directions, scripts, and the most recent FreshTrain releases are available at github.com/McMahonLab/TaxAss.

Several processed datasets have been produced from TYMEFLIES and are available publicly. Quality filtered metagenome read files (fastq files) and single-sample assemblies were created

by the JGI IMG/M pipeline and are available from their website (https://img.jgi.doe.gov) under ITS Proposal ID 504350. A set 2,800 of MAGs created by cross-mapping samples to single-sample assemblies (Mattock and Watson 2023) and de-replicating to 96 percent average nucleotide identity (Olm et al. 2017) was created by Rohwer *et al.* (2024a) and is available from NCBI under BioProject accession number PRJNA1158976. A co-assembly of all metagenome samples was also performed by JGI and is available under NCBI BioProject accession PRJNA1134257, as well as 1,800 accompanying MAGs that are available from IMG/M under taxon identifier 3300059473 (Oliver et al. 2024).

Available Environmental Data

Lake Mendota is part of the NTL-LTER site, which conducts routine sampling of water chemistry, lake physical characteristics, phytoplankton and zooplankton, macrophytes, and fish. These relevant datasets span the time period of the limony dataset and are all available through the Environmental Data Initiative. Users can find all available data by searching for datasets associated with the NTL site. In particular, users may be interested in the available long-term records of water temperature (Robertson 2016; Magnuson et al. 2020, 2021a; b; Rohwer and McMahon 2022a), ice cover (Robertson 2016; Magnuson et al. 2021c), water clarity (Robertson 2016; Magnuson et al. 2021d; Rohwer and McMahon 2022b), dissolved oxygen (Magnuson et al. 2021a; b; Rohwer and McMahon 2022a), and various nutrients and ions including carbon, nitrogen, and phosphorus (Magnuson et al. 2023a; b).

Using the limony R package.

The processed limony data is available as part of an R package that can be installed using the command `devtools::install_github("rrohwer/limony")`. The R package is hosted at github.com/rrohwer/limony. In addition to the processed data, the R package includes scripts to easily subset the dataset by sample date and taxonomy, while maintaining the nested list data structure.

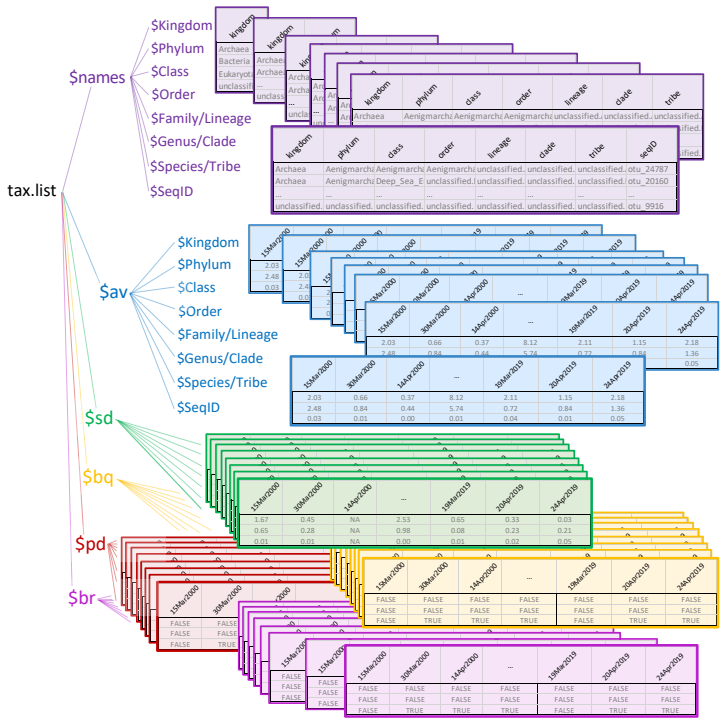

Fig 3. Limony nested list data structure. A visual representation of the nested list data structure in which processed data are provided as part of the limony R package. Abbreviations: av = average, sd = standard deviation, bq = below quantification, pd = partially detected, br = below resolution.

Nested list data structure. The processed limony data includes the taxonomic assignments, the average abundances, and the standard deviations of each OTU, as well as information about whether it was confidently observed. This information is saved in a list of matrices with

matching dimensions named “av” (average), “sd” (standard deviation), “bq” (below quantification), “pd” (partial detection), and “br” (below resolution). The elements bq, pd, and br are boolean matrices, and bq is TRUE if either pd or br is TRUE. A visual schematic of this data structure is shown in Fig 3. Nested under each categorical element is a list corresponding to each taxonomic level. These levels are Kingdom, Phylum, Class, Order, Family/Lineage, Genus/Clade, Species/Tribe, and SeqID. Each level coarser than SeqID has been grouped, so that all categorical elements have been recalculated. The average of a taxon group is the sum of all daughter group averages, and the standard deviation is the propagated error of sums of the daughter standard deviations. The boolean elements bq, pd, and br are TRUE if all of their daughter constituents are TRUE.

Subsetting the nested list. The limony package includes functions to subset samples and taxa while maintaining the nested list structure. For example, users can subset particular sample dates, and they can exclude non-standard samples from analysis. Additionally, users can subset a particular taxonomic group of interest. Alternately, If users prefer to work with a simple OTU table, this can be obtained by subsetting the list. Examples of all of these workflows are available in the limony R package vignette, which can be viewed without installation on the limony GitHub repository (github.com/rrohwer/limony). In addition to the vignette examples, documentation is provided for all limony package functions.

Code Availability

Code used to create the processed limony data is available at github.com/rrohwer/limony along with the limony R package.

Acknowledgements

The maintenance of a long-term observatory requires continuous support from numerous personnel, including professors, postdoctoral fellows, technicians, graduate students, and undergraduate students. We thank all of these researchers, including sampling leads: Angela Kent, Tony Yannarell, Ashley Shade, Stuart Jones, Ryan Newton, Georgia Wolfe, Emily Kara Read, Todd Miller, Lucas Beversdorf, James Mutschler, and Riley Hale. We also thank the initial Microbial Observatory PI Eric W. Triplett.

Maintaining a long-term observational study is not possible without continuous funding. We thank all of these funding agencies for making this dataset possible:

- U.S. National Science Foundation Postdoctoral Research Fellowship in Biology, award number 2011002.
- E. Michael and Winona Foster WARF Wisconsin Idea Fellowship, 2018.
- National Institute of Food and Agriculture, U.S. Department of Agriculture, award number 2016-67012-24709.
- National Institute of Food and Agriculture, U.S. Department of Agriculture, Hatch Projects WIS01516, WIS01789, WIS03004.
- U.S. National Science Foundation North Temperate Lakes Long-Term Ecological Research site, NTL-LTER, award numbers DEB-9632853, DEB-0217533, DEB-0822700, and DEB-1440297.
- U.S. National Science Foundation Microbial Observatories program, award numbers MCB-9977903 and DEB-0702395.
- U.S. National Science Foundation INSPIRE award, DEB-1344254.

Finally, we thank Joshua J. Hamilton for securing funding for amplicon sequencing (U.S. Department of Agriculture award number 2016-67012-24709) and the DOE Joint Genome

Institute for metagenome sequencing through CSP 504350. We also thank the Fuhrman Lab for generously supplying their mock communities.

Author contributions

- RRR maintained the fieldwork from 2013 – 2019, extracted the DNA, applied quality control and pipeline processing, wrote the scripts for custom data processing, created the R package, created the figures, and wrote the manuscript.
- KDM maintained the fieldwork from 2003 – 2019, advised the sample preparation and data processing, secured funding, and edited the manuscript.

Competing interests

The authors declare no competing interests.

Supplemental Materials

Online-only Table 1. Sample metadata. This table links sample names with limony and TYMEFLIES sample IDs and NCBI SRA accessions, as well as with metadata such as sample dates and notes about any non-standard locations, depths, or filtration.

References

Amir, A. and others. 2017. Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns. mSystems **2**. doi:10.1128/mSystems.00191-16

Beversdorf, L. J., T. R. Miller, and K. D. McMahon. 2015. Long-term monitoring reveals carbon–nitrogen metabolism key to microcystin production in eutrophic lakes. Front. Microbiol. **6**. doi:10.3389/fmicb.2015.00456

Beversdorf, L., C. Weirich, S. Bartlett, and Todd. Miller. 2017. Variable Cyanobacterial Toxin and Metabolite Profiles across Six Eutrophic Lakes of Differing Physiochemical Characteristics. Toxins **9**: 62. doi:10.3390/toxins9020062

Brock, T. D. 1985. A eutrophic lake: Lake Mendota, Wisconsin, Springer-Verlag.

Bunse, C., and J. Pinhassi. 2017. Marine Bacterioplankton Seasonal Succession Dynamics. Trends in Microbiology **25**: 494–505. doi:10.1016/j.tim.2016.12.013

Callahan, B. J., P. J. McMurdie, M. J. Rosen, A. W. Han, A. J. A. Johnson, and S. P. Holmes. 2016. DADA2: High-resolution sample inference from Illumina amplicon data. Nat Meth **13**: 581–583. doi:10.1038/nmeth.3869

469 Carey, C. C., P. C. Hanson, R. C. Lathrop, S. Amand, and A. L. 2016. Using wavelet analyses to
470 examine variability in phytoplankton seasonal succession and annual periodicity. J
471 Plankton Res **38**: 27–40. doi:10.1093/plankt/fbv116
472 Carpenter, S. R., E. G. Booth, C. J. Kucharik, and R. C. Lathrop. 2015. Extreme daily loads: role
473 in annual phosphorus input to a north temperate lake. Aquat Sci **77**: 71–79.
474 doi:10.1007/s00027-014-0364-5
475 D'Amore, R. and others. 2016. A comprehensive benchmarking study of protocols and
476 sequencing platforms for 16S rRNA community profiling. BMC Genomics **17**: 55.
477 doi:10.1186/s12864-015-2194-9
478 Edgar, R. C. 2016. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon
479 sequencing. bioRxiv 081257. doi:10.1101/081257
480 Eren, A. M., H. G. Morrison, P. J. Lescault, J. Reveillaud, J. H. Vineis, and M. L. Sogin. 2015.
481 Minimum entropy decomposition: Unsupervised oligotyping for sensitive partitioning
482 of high-throughput marker gene sequences. The ISME Journal **9**: 968–979.
483 doi:10.1038/ismej.2014.195
484 Estaki, M. and others. 2020. QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse
485 Microbiome Data and Comparative Studies with Publicly Available Data. Current
486 Protocols in Bioinformatics **70**: e100. doi:10.1002/cpbi.100
487 Fisher, M. M., and E. W. Triplett. 1999. Automated Approach for Ribosomal Intergenic Spacer
488 Analysis of Microbial Diversity and Its Application to Freshwater Bacterial
489 Communities. Appl. Environ. Microbiol. **65**: 4630–4636.
490 Hall, M. W., R. R. Rohwer, J. Perrie, K. D. McMahon, and R. G. Beiko. 2017. Ananke: temporal
491 clustering reveals ecological dynamics of microbial communities. PeerJ **5**: e3812.
492 doi:10.7717/peerj.3812
493 Herren, C. M., and K. D. McMahon. 2017. Cohesion: a method for quantifying the connectivity
494 of microbial communities. The ISME Journal **11**: 2426. doi:10.1038/ismej.2017.91

Kane, M. D. 2004. Microbial Observatories: Exploring and Discovering Microbial Diversity in the 21st Century. Microb Ecol **48**: 447–448. doi:10.1007/s00248-004-9953-5

Kara, E. L., P. C. Hanson, Y. H. Hu, L. Winslow, and K. D. McMahon. 2013. A decade of seasonal dynamics and co-occurrences within freshwater bacterioplankton communities from eutrophic Lake Mendota, WI, USA. The ISME Journal **7**: 680. doi:10.1038/ismej.2012.118

Krinos, A. I., R. M. Bowers, R. R. Rohwer, K. D. McMahon, T. Woyke, and F. Schulz. 2024. Time-series metagenomics reveals changing protistan ecology of a temperate dimictic lake. Microbiome **12**: 133. doi:10.1186/s40168-024-01831-y

Ladwig, R., A. P. Appling, A. Delany, H. A. Dugan, Q. Gao, N. Lottig, J. Stachelek, and P. C. Hanson. 2022. Long-term change in metabolism phenology in north temperate lakes. Limnology and Oceanography **67**: 1502–1521. doi:10.1002/lno.12098

Ladwig, R., P. C. Hanson, H. A. Dugan, C. C. Carey, Y. Zhang, L. Shu, C. J. Duffy, and K. M. Cobourn. 2021. Lake thermal structure drives interannual variability in summer anoxia dynamics in a eutrophic lake over 37 years. Hydrology and Earth System Sciences **25**: 1009–1032. doi:10.5194/hess-25-1009-2021

Larsson, A. J. M., G. Stanley, R. Sinha, I. L. Weissman, and R. Sandberg. 2018. Computational correction of index switching in multiplexed sequencing libraries. Nature Methods **15**: 305–307. doi:10.1038/nmeth.4666

Lathrop, R. C. 2007. Perspectives on the eutrophication of the Yahara lakes. Lake and Reservoir Management **23**: 345–365. doi:10.1080/07438140709354023

Lathrop, R. C., and S. R. Carpenter. 2014. Water quality implications from three decades of phosphorus loads and trophic dynamics in the Yahara chain of lakes. Inland Waters **4**: 1–14. doi:10.5268/IW-4.1.680

Lathrop, R. C., S. R. Carpenter, and L. G. Rudstam. 1996. Water clarity in Lake Mendota since 1900: responses to differing levels of nutrients and herbivory. Canadian Journal of Fisheries and Aquatic Sciences **53**: 2250–2261.

Lennon, J. T. 2011. Replication, lies and lesser-known truths regarding experimental design in environmental microbiology. Environmental Microbiology **13**: 1383–1386. doi:10.1111/j.1462-2920.2011.02445.x

Magee, M. R., P. B. McIntyre, P. C. Hanson, and C. H. Wu. 2019. Drivers and Management Implications of Long-Term Cisco Oxythermal Habitat Decline in Lake Mendota, WI. Environmental Management **63**: 396–407. doi:10.1007/s00267-018-01134-7

Magee, M. R., C. H. Wu, D. M. Robertson, R. C. Lathrop, and D. P. Hamilton. 2016. Trends and abrupt changes in 104 years of ice cover and water temperature in a dimictic lake in response to air temperature, wind speed, and water clarity drivers. Hydrology and Earth System Sciences **20**: 1681–1702. doi:10.5194/hess-20-1681-2016

Magnuson, J., S. Carpenter, and E. Stanley. 2020. North Temperate Lakes LTER: High Frequency Water Temperature Data - Lake Mendota Buoy 2006 - current.doi:10.6073/PASTA/8CEFF296AD68FA8DA6787076E0A5D992

Magnuson, J. J., S. R. Carpenter, and E. H. Stanley. 2021a. North Temperate Lakes LTER: Physical Limnology of Primary Study Lakes 1981 - current.doi:10.6073/PASTA/316203040EA1B8ECE89673985AB431B7

Magnuson, J. J., S. R. Carpenter, and E. H. Stanley. 2021b. Lake Mendota Multiparameter Sonde Profiles: 2017 - current.doi:10.6073/PASTA/5F15BF453851987FC030B2F07A110B21

Magnuson, J. J., S. R. Carpenter, and E. H. Stanley. 2021c. North Temperate Lakes LTER: Ice Duration - Madison Lakes Area 1853 - current.doi:10.6073/PASTA/AB31F2489EE436BEB73FC8F1D0213D97

Magnuson, J. J., S. R. Carpenter, and E. H. Stanley. 2021d. North Temperate Lakes LTER: Secchi Disk Depth; Other Auxiliary Base Crew Sample Data 1981 - current.doi:10.6073/PASTA/26FA98B39F9758FDA2109021F5B88076

Magnuson, J. J., S. R. Carpenter, and E. H. Stanley. 2023a. North Temperate Lakes LTER: Chemical Limnology of Primary Study Lakes: Nutrients, pH and Carbon 1981 - current.doi:10.6073/PASTA/325232E6E4CD1CE04025FA5674F7B782

Magnuson, J. J., S. R. Carpenter, and E. H. Stanley. 2023b. North Temperate Lakes LTER: Chemical Limnology of Primary Study Lakes: Major Ions 1981 - current.doi:10.6073/pasta/bb563f16c7338fdb3ddf82057ef43cc6

Matsuzaki, S.-I. S., R. C. Lathrop, S. R. Carpenter, J. R. Walsh, M. J. V. Zanden, M. R. Gahler, and E. H. Stanley. 2020. Climate and food web effects on the spring clear-water phase in two north-temperate eutrophic lakes. Limnol. Oceanogr. **n/a**: 1–17. doi:10.1002/lno.11584

Mattock, J., and M. Watson. 2023. A comparison of single-coverage and multi-coverage metagenomic binning reveals extensive hidden contamination. Nat Methods **20**: 1170–1173. doi:10.1038/s41592-023-01934-8

McMurdie, P. J., and S. Holmes. 2014. Waste Not, Want Not: Why Rarefying Microbiome Data Is Inadmissible. PLOS Computational Biology **10**: e1003531. doi:10.1371/journal.pcbi.1003531

Nelson, M. C., H. G. Morrison, J. Benjamino, S. L. Grim, and J. Graf. 2014. Analysis, Optimization and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. PLOS ONE **9**: e94249. doi:10.1371/journal.pone.0094249

Newton, R. J., S. E. Jones, A. Eiler, K. D. McMahon, and S. Bertilsson. 2011. A Guide to the Natural History of Freshwater Lake Bacteria. Microbiology and Molecular Biology Reviews **75**: 14–49. doi:10.1128/MMBR.00028-10

569 Oliver, T. and others. 2024. Coassembly and binning of a twenty-year metagenomic time-series
570 from Lake Mendota. Sci Data **11**: 966. doi:10.1038/s41597-024-03826-8
571 Olm, M. R., C. T. Brown, B. Brooks, and J. F. Banfield. 2017. dRep: a tool for fast and accurate
572 genomic comparisons that enables improved genome recovery from metagenomes
573 through de-replication. ISME J **11**: 2864–2868. doi:10.1038/ismej.2017.126
574 Prosser, J. I. 2010. Replicate or lie. Environmental Microbiology **12**: 1806–1810.
575 doi:10.1111/j.1462-2920.2010.02201.x
576 Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner.
577 2013. The SILVA ribosomal RNA gene database project: improved data processing and
578 web-based tools. Nucleic Acids Research **41**: D590–D596. doi:10.1093/nar/gks1219
579 Robertson, D. 2016. Lake Mendota water temperature secchi depth snow depth ice thickness
580 and meterological conditions 1894 -
581 2007.doi:10.6073/PASTA/F20F9A644BD12E4B80CB288F1812C935
582 Rognes, T., T. Flouri, B. Nichols, C. Quince, and F. Mahé. 2016. VSEARCH: a versatile open
583 source tool for metagenomics. PeerJ **4**: e2584. doi:10.7717/peerj.2584
584 Rohwer, R. R., R. J. Hale, M. J. Vander Zanden, T. R. Miller, and K. D. McMahon. 2023. Species
585 invasions shift microbial phenology in a two-decade freshwater time series.
586 Proceedings of the National Academy of Sciences **120**: e2211796120.
587 doi:10.1073/pnas.2211796120
588 Rohwer, R. R., J. J. Hamilton, R. J. Newton, and K. D. McMahon. 2018. TaxAss: Leveraging a
589 Custom Freshwater Database Achieves Fine-Scale Taxonomic Resolution. mSphere **3**:
590 e00327-18. doi:10.1128/mSphere.00327-18
591 Rohwer, R. R., M. Kirkpatrick, S. L. Garcia, M. Kellom, K. D. McMahon, and B. J. Baker. 2024a.
592 Bacterial ecology and evolution converge on seasonal and decadal scales.
593 2024.02.06.579087. doi:10.1101/2024.02.06.579087

Rohwer, R. R., R. Ladwig, P. C. Hanson, J. R. Walsh, M. J. Vander Zanden, and H. A. Dugan. 2024b. Increased anoxia following species invasion of a eutrophic lake. Limnology and Oceanography Letters **9**: 33–42. doi:10.1002/lol2.10364

Rohwer, R. R., and K. D. McMahon. 2022a. Lake Mendota Microbial Observatory Temperature, Dissolved Oxygen, pH, and conductivity data, 2006-present.doi:10.6073/PASTA/7E533C197ED8EBD27777A89A2C8D7DFE

Rohwer, R. R., and K. D. McMahon. 2022b. Lake Mendota Microbial Observatory Secchi Disk Measurements 2012-present.doi:10.6073/PASTA/3B650E19D28CBC7B9ED631F0A7878033

Schloss, P. D. 2020. Reintroducing mothur: 10 Years Later. Appl. Environ. Microbiol. **86**. doi:10.1128/AEM.02343-19

Shade, A., A. D. Kent, S. E. Jones, R. J. Newton, E. W. Triplett, and K. D. McMahon. 2007. Interannual dynamics and phenology of bacterial communities in a eutrophic lake. Limnol. Oceanogr. **52**: 487–494. doi:10.4319/lo.2007.52.2.0487

Spear, M. J., P. A. Wakker, T. P. Shannon, R. L. Lowe, L. E. Burlakova, A. Y. Karatayev, and M. J. Vander Zanden. 2022. Early changes in the benthic community of a eutrophic lake following zebra mussel (Dreissena polymorpha) invasion. Inland Waters **0**: 1–19. doi:10.1080/20442041.2021.2007744

van der Valk, T., F. Vezzi, M. Ormestad, L. Dalen, and K. Guschanski. 2018. Index hopping on the Illumina HiseqX platform and its consequences for ancient DNA studies. bioRxiv. doi:10.1101/179028

Vander Zanden, M. J. and others. 2024. Nine Lessons about Aquatic Invasive Species from the North Temperate Lakes Long-Term Ecological Research (NTL-LTER) Program. BioScience **74**: 509–523. doi:10.1093/biosci/biae062

618 Walsh, J. R., S. R. Carpenter, and M. J. V. Zanden. 2016a. Invasive species triggers a massive
619 loss of ecosystem services through a trophic cascade. PNAS **113**: 4081–4085.
620 doi:10.1073/pnas.1600366113
621 Walsh, J. R., R. C. Lathrop, and M. J. Vander Zanden. 2017. Invasive invertebrate predator,
622 Bythotrephes longimanus, reverses trophic cascade in a north-temperate lake. Limnol.
623 Oceanogr. **62**: 2498–2509. doi:10.1002/lno.10582
624 Walsh, J. R., S. E. Munoz, and M. J. Vander Zanden. 2016b. Outbreak of an undetected invasive
625 species triggered by a climate anomaly. Ecosphere **7**: n/a-n/a. doi:10.1002/ecs2.1628
626 Walsh, J. R., M. J. Spear, T. P. Shannon, P. J. Krysan, and M. J. Vander Zanden. 2019. Using
627 eDNA, sediment subfossils, and zooplankton nets to detect invasive spiny water flea
628 (Bythotrephes longimanus). Biol Invasions **21**: 377–389. doi:10.1007/s10530-018-
629 1862-5
630 Walters, W. and others. 2016. Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal
631 Internal Transcribed Spacer Marker Gene Primers for Microbial Community Surveys.
632 mSystems **1**: e00009-15. doi:10.1128/mSystems.00009-15
633 White, E. R. 2019. Minimum Time Required to Detect Population Trends: The Need for Long-
634 Term Monitoring Programs. BioScience **69**: 40–46. doi:10.1093/biosci/biy144
635 Wright, E. S., and K. H. Vetsigian. 2016. Quality filtering of Illumina index reads mitigates
636 sample cross-talk. BMC Genomics **17**: 876. doi:10.1186/s12864-016-3217-x
637 Yannarell, A. C., A. D. Kent, G. H. Lauster, T. K. Kratz, and E. W. Triplett. 2003. Temporal
638 Patterns in Bacterial Communities in Three Temperate Lakes of Different Trophic
639 Status. Microb Ecol **46**: 391–405. doi:10.1007/s00248-003-1008-9
640 Yeh, Y.-C., D. M. Needham, E. T. Sieradzki, and J. A. Fuhrman. 2018. Taxon Disappearance from
641 Microbiome Analysis Reinforces the Value of Mock Communities as a Standard in
642 Every Sequencing Run. mSystems **3**: 9.

Zhou, Z., P. Q. Tran, C. Martin, R. R. Rohwer, B. J. Baker, K. D. McMahon, and K. Anantharaman. 2024. A 20-year time-series of a freshwater lake reveals seasonal dynamics and environmental drivers of viral diversity, ecology, and evolution. 2024.02.06.579183. doi:10.1101/2024.02.06.579183