Lake iTag measurements over nineteen years, introducing the limony dataset

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

corresponding author(s): Robin Rohwer (robin.rohwer@gmail.com)

Abstract

Lakes are focal points of biogeochemical cycling on the terrestrial landscape, and microbial communities are the machines driving these processes. However, our view into how freshwater microbial communities function is obscured by high short-term variability. We provide a long-term, two-decade time series (years 2000 – 2019) of a freshwater microbial community characterized with 16S rRNA gene amplicon sequencing. We provide raw sequencing data from 1,023 total sequenced samples and controls, as well as references to paired publicly available environmental data. Moreover, processed, curated, and formatted microbial data are also available for easy download as an R package, along with well-documented functions for subsetting the data by samples and taxa. This processed dataset includes 970 total replicates from 496 samples. The "limony" dataset and its paired environmental data is a complete and ready-to-use resource for ecologists, modellers, educators, data scientists, and others from diverse fields.

Background & Summary

Lake Mendota

Lake Mendota is in southern Wisconsin (USA), on the shores of the University of Wisconsin-Madison and the Wisconsin state capitol (Fig 1). Lake Mendota is dimictic and temperate, meaning that the water column mixes vertically twice a year 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 [1,2]. Vertical mixing occurs immediately after ice off in April, and in October [3–7].

Summer surface water temperatures typically reach 25°C, and the lake is popular for recreational activities such as fishing, sailing, and swimming. Lake Mendota is eutrophic, meaning that it has high concentrations of nitrogen and phosphorus and is plagued by harmful cyanobacterial blooms in the summer [8]. 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 [3,9–12]. Summer water clarities of 1-2 m are punctuated by surface Cyanobacterial blooms, which prevent recreation due to cyanotoxins and noxious odors [13,14]. Lake Mendota is considered to be a fairly typical temperate eutrophic 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 [15,16].

Long-term research on Lake Mendota (WI, USA) began in the early 1900s [13,15] 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)[17] 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". As a result of this intensive study, a wealth of long-term environmental data is publicly available through the Environmental Data Initiative (EDI).

The lake iTag measurements over nineteen years dataset, known by the acronym limony, spans the duration at the time of sequencing of the archived Lake Mendota Microbial Observatory, from 2000-2019 (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, and increased phosphorus loading from heavy rains [1,10,18,19]. However, during the two decades covered by the limony dataset Lake Mendota has also experienced several abrupt ecological shifts. 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 [20–22]. This major food web disturbance is not unrelated to climate, because an unusually cool July enabled the population explosion, even though spiny water flea had existed at low levels in Lake Mendota since the 1990s [23]. Another sudden shift to Lake Mendota's food

web occurred with the invasion of zebra mussels (*Dreissena polymorpha*), which were first detected in the fall of 2015. By the summer of 2016 these 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. Zebra mussels became widespread enough to theoretically filter the epilimnion volume every 45 days, but the water clarity of the lake did not change[24]. The impacts of these species invasions on the microbial community are further explored in Rohwer *et al.* (*submitted*) [25].

980 iTags from 477 Sample Dates spanning 19 Years

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 [26]. Regardless of generation times, multi-decadal time series are needed to tease apart true long-term change from large underlying multi-scale variability [27].

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[28–32] it. 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 addressed by different data processing pipelines, and the environmental noise, which is typically ignored in microbial ecology [33,34].

Here we describe in detail the methods used in collecting and sequencing samples, the pipeline used to create available processed data, and additional resources that complement limony. We hope this background and practical guidance will enable other scientists to analyse limony in a way that is conscious of its limitations while taking advantage of its strengths.

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 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 0.2 μm 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 0.2 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 and Sequencing

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. Each group was spread over two sequencing plates, so that duplicates were sequenced in separate Illumina runs.

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. The first and last extraction of every 100-extraction kit was performed on a sterile filter as a blank. All

133 extractions were performed by the same person according to the same method, and no
134 differences in DNA yield were observed by extraction date or kit.

135 **Amplification and Sequencing.** 16S rRNA gene amplicon sequencing was performed
136 at Argonne National Lab according the procedure established by Walters *et al.* [35] and
137 recommended by the Earth Microbiome Project. The V4-V5 variable region was amplified
138 (515F-Y GTGYCAGCMGCCGCGGTAA, 926R CCGYCAATTYMTTTRAGTTT), and sequencing was
139 performed on an Illumina MiSeq instrument using version 2 chemistry. Sequencing was split
140 across 4 plates that multiplexed 245, 247, 269, and 262 samples each.

141

142 Data Processing

143 **Quality filtering and demultiplexing.** Before demultiplexing, index reads were quality
144 filtered using vsearch [36] (vsearch v2.13.3) to retain only index reads where the expected errors
145 were ≤ 0.03 (`--fastq_maxee = 0.03`). Mothur was then used to remove the matching
146 sequences for the forward and reverse read files [32] (mothur v.1.44.1) (`list.seqs()` and
147 `get.seqs()`). Quality control of index reads can remove cross-contamination that occurs
148 during sequencing, sometimes called cross-talk or index hopping. This is not typically
149 considered a major concern with the MiSeq platform [37,38], but it has been observed at low
150 levels ranging from 0.16 - 0.2% [39,40]. We chose a maximum expected errors cut-off of 0.03, to
151 approximate the mean quality score cutoff of 25 suggested by Wright and Vestigian [41].
152 Read quality cut-offs were chosen for each of the four limony sequencing plates separately.
153 Cut-offs for Read quality were chosen by running vsearch [36] with a range of expected error
154 cutoffs and truncation lengths (`--fastq_trunclen`), and choosing a cutoff of acceptable
155 quality and length that retained the most sequences. A maximum expected errors of 2 (`--`
156 `fastq_maxee = 2`) was applied to all forward and reverse reads following truncation of the
157 forward reads at 247 bp (plates 1 and 2) or 250 bp (plates 3 and 4) and truncation of the
158 reverse reads at 195 bp (plates 1 and 2) or 192 bp (plates 3 and 4). Sequences were
159 demultiplexed using qiime [42] (qiime2-2019.4) without golay correction of the barcode
160 sequences.

161 **Error correction using dada2.** For dada2 error correction (dada2 version 1.12.1) [29],
162 learning errors and pooled error correction were run separately on each plate, as
163 recommended by the dada2 big data workflow tutorial. After removing phiX and merging
164 forward and reverse reads, the four plates were combined for chimera detection.

165 **Taxonomy assignment using TaxAss with the FreshTrain and Silva.** We assigned
166 taxonomy with the Silva database [43] (silva138) and the FreshTrain database [44]

(FreshTrain15Jun2020silva138) using the TaxAss workflow [45] (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).

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 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 [46].

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.

Data Records

Raw sequencing data is available from the NCBI Sequence Read Archive under accession number PRJNA846788 (Reviewer Link: https://dataview.ncbi.nlm.nih.gov/object/PRJNA846788?reviewer=varcj4a5hng2m7n1bcde77oecg). A key linking sequencing codes to sample names 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. Environmental data collected at the same time as the microbial samples are 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 [7,12]. 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 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 MiSeq plates, and can be used to detect kit contamination, cross contamination, and sequencing cross-talk. Each of the four 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.* [47]. The two mock communities were an even and a staggered marine mock community provided by the Fuhrman Lab [47]. 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.

Usage Notes

Sample Metadata

A table that lists sample names, sample dates, sequencing codes (as identified on 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 [44]. This Freshwater Training set (the FreshTrain) enables fine-resolution taxonomy classifications when used with a Taxonomy Assignment tool (TaxAss) to classify 16S rRNA gene amplicon data [45]. 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.

The microbial community of Lake Mendota has been extensively studied prior to the limony dataset, 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 [48]. These ARISA-based studies found consistent seasonal changes in bacterial community composition [49–51]. 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 [52,53].

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-LTER site.

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.

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 is available at github.com/rrohwer/limony. This repository hosts the scripts used to process the limony data, as well as an R package that includes the processed limony data.

Acknowledgements

The maintenance of a long-term observatory requires continuous support from numerous personnel, including professors, 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 Muetschler, and Riley Hale. We also thank the initial Microbial Observatory PI Eric W. Triplett. Finally, we thank Joshua J. Hamilton for securing funding for amplicon sequencing.

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.

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 data processing, secured funding, and edited the manuscript.

Competing interests

The authors declare no competing interests.

References

1. Magee, M. R., Wu, C. H., Robertson, D. M., Lathrop, R. C. & Hamilton, D. P. 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 (2016).
2. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North Temperate Lakes LTER: Ice Duration - Madison Lakes Area 1853 - current. *Environmental Data Initiative* **knb-lter-ntl.33**, (2021).
3. Robertson, D. M. Lake Mendota water temperature secchi depth snow depth ice thickness and meterological conditions 1894 - 2007. *Environmental Data Initiative* **knb-lter-ntl.335**, (2016).
4. Magnuson, J., Carpenter, S. & Stanley, E. North Temperate Lakes LTER: High Frequency Water Temperature Data - Lake Mendota Buoy 2006 - current. *Environmental Data Initiative* **knb-lter-ntl.130**, (2020).
5. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North Temperate Lakes LTER: Physical Limnology of Primary Study Lakes 1981 - current. *Environmental Data Initiative* **knb-lter-ntl.29**, (2021).
6. Ladwig, R. *et al.* 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 (2021).
7. Rohwer, R. R. & McMahon, K. D. Lake Mendota Microbial Observatory Temperature, Dissolved Oxygen, pH, and conductivity data, 2006-present. *Environmental Data Initiative* **knb-lter-ntl.415**, (2022).
8. Lathrop, R. C. Perspectives on the eutrophication of the Yahara lakes. *Lake and Reservoir Management* **23**, 345–365 (2007).

9. Carey, C. C., Hanson, P. C., Lathrop, R. C., Amand, S. & L, A. Using wavelet analyses to examine variability in phytoplankton seasonal succession and annual periodicity. *J Plankton Res* **38**, 27–40 (2016).
10. Matsuzaki, S.-I. S. *et al.* Climate and food web effects on the spring clear-water phase in two north-temperate eutrophic lakes. *Limnol. Oceanogr.* **n/a**, 1–17 (2020).
11. Magnuson, J. J., Carpenter, S. R. & Stanley, E. H. North Temperate Lakes LTER: Secchi Disk Depth; Other Auxiliary Base Crew Sample Data 1981 - current. *Environmental Data Initiative* **knb-lter-ntl.31**, (2021).
12. Rohwer, R. R. & McMahon, K. D. Lake Mendota Microbial Observatory Secchi Disk Measurements 2012-present. *Environmental Data Initiative* **knb-lter-ntl.416**, (2022).
13. Lathrop, R. C., Carpenter, S. R. & Rudstam, L. G. 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 (1996).
14. Beversdorf, L., Weirich, C., Bartlett, S. & Miller, Todd. Variable Cyanobacterial Toxin and Metabolite Profiles across Six Eutrophic Lakes of Differing Physiochemical Characteristics. *Toxins* **9**, 62 (2017).
15. Brock, T. D. *A eutrophic lake: Lake Mendota, Wisconsin*. (Springer-Verlag, 1985).
16. Lathrop, R. C. & Carpenter, S. R. Water quality implications from three decades of phosphorus loads and trophic dynamics in the Yahara chain of lakes. *Inland Waters* **4**, 1–14 (2014).
17. Kane, M. D. Microbial Observatories: Exploring and Discovering Microbial Diversity in the 21st Century. *Microb Ecol* **48**, 447–448 (2004).
18. Carpenter, S. R., Booth, E. G., Kucharik, C. J. & Lathrop, R. C. Extreme daily loads: role in annual phosphorus input to a north temperate lake. *Aquat Sci* **77**, 71–79 (2015).

369 19. Magee, M. R., McIntyre, P. B., Hanson, P. C. & Wu, C. H. Drivers and Management
370 Implications of Long-Term Cisco Oxythermal Habitat Decline in Lake Mendota, WI.
371 *Environmental Management* **63**, 396–407 (2019).
372 20. Walsh, J. R., Carpenter, S. R. & Zanden, M. J. V. Invasive species triggers a massive loss of
373 ecosystem services through a trophic cascade. *PNAS* **113**, 4081–4085 (2016).
374 21. Walsh, J. R., Lathrop, R. C. & Vander Zanden, M. J. Invasive invertebrate predator,
375 Bythotrephes longimanus, reverses trophic cascade in a north-temperate lake. *Limnol.*
376 *Oceanogr.* **62**, 2498–2509 (2017).
377 22. Walsh, J. R., Spear, M. J., Shannon, T. P., Krysan, P. J. & Vander Zanden, M. J. Using eDNA,
378 sediment subfossils, and zooplankton nets to detect invasive spiny water flea
379 (Bythotrephes longimanus). *Biol Invasions* **21**, 377–389 (2019).
380 23. Walsh, J. R., Munoz, S. E. & Vander Zanden, M. J. Outbreak of an undetected invasive
381 species triggered by a climate anomaly. *Ecosphere* **7**, n/a-n/a (2016).
382 24. Spear, M. J. *et al.* Early changes in the benthic community of a eutrophic lake following
383 zebra mussel (Dreissena polymorpha) invasion. *Inland Waters* **0**, 1–19 (2022).
384 25. Rohwer, R. R., Hale, R. J., Miller, T. R. & McMahon, K. D. Species invasions shift microbial
385 phenology in a two-decade freshwater time series. *in prep.* (in prep).
386 26. Bunse, C. & Pinhassi, J. Marine Bacterioplankton Seasonal Succession Dynamics. *Trends in*
387 *Microbiology* **25**, 494–505 (2017).
388 27. White, E. R. Minimum Time Required to Detect Population Trends: The Need for Long-
389 Term Monitoring Programs. *BioScience* **69**, 40–46 (2019).
390 28. Eren, A. M. *et al.* Minimum entropy decomposition: Unsupervised oligotyping for sensitive
391 partitioning of high-throughput marker gene sequences. *The ISME Journal* **9**, 968–979
392 (2015).
393 29. Callahan, B. J. *et al.* DADA2: High-resolution sample inference from Illumina amplicon
394 data. *Nat Meth* **13**, 581–583 (2016).

395 30. Edgar, R. C. UNOISE2: improved error-correction for Illumina 16S and ITS amplicon
396 sequencing. *bioRxiv* 081257 (2016) doi:10.1101/081257.
397 31. Amir, A. *et al.* Deblur Rapidly Resolves Single-Nucleotide Community Sequence Patterns.
398 *mSystems* **2**, (2017).
399 32. Schloss, P. D. Reintroducing mothur: 10 Years Later. *Appl. Environ. Microbiol.* **86**, (2020).
400 33. Prosser, J. I. Replicate or lie. *Environmental Microbiology* **12**, 1806–1810 (2010).
401 34. Lennon, J. T. Replication, lies and lesser-known truths regarding experimental design in
402 environmental microbiology. *Environmental Microbiology* **13**, 1383–1386 (2011).
403 35. Walters, W. *et al.* Improved Bacterial 16S rRNA Gene (V4 and V4-5) and Fungal Internal
404 Transcribed Spacer Marker Gene Primers for Microbial Community Surveys. *mSystems* **1**,
405 e00009-15 (2016).
406 36. Rognes, T., Flouri, T., Nichols, B., Quince, C. & Mahé, F. VSEARCH: a versatile open source
407 tool for metagenomics. *PeerJ* **4**, e2584 (2016).
408 37. Larsson, A. J. M., Stanley, G., Sinha, R., Weissman, I. L. & Sandberg, R. Computational
409 correction of index switching in multiplexed sequencing libraries. *Nature Methods* **15**,
410 305–307 (2018).
411 38. van der Valk, T., Vezzi, F., Ormestad, M., Dalen, L. & Guschanski, K. Index hopping on the
412 Illumina HiseqX platform and its consequences for ancient DNA studies. *bioRxiv* (2018)
413 doi:10.1101/179028.
414 39. Nelson, M. C., Morrison, H. G., Benjamino, J., Grim, S. L. & Graf, J. Analysis, Optimization
415 and Verification of Illumina-Generated 16S rRNA Gene Amplicon Surveys. *PLOS ONE* **9**,
416 e94249 (2014).
417 40. D'Amore, R. *et al.* A comprehensive benchmarking study of protocols and sequencing
418 platforms for 16S rRNA community profiling. *BMC Genomics* **17**, 55 (2016).
419 41. Wright, E. S. & Vetsigian, K. H. Quality filtering of Illumina index reads mitigates sample
420 cross-talk. *BMC Genomics* **17**, 876 (2016).

421 42. Estaki, M. *et al.* QIIME 2 Enables Comprehensive End-to-End Analysis of Diverse
422 Microbiome Data and Comparative Studies with Publicly Available Data. *Current Protocols*
423 *in Bioinformatics* **70**, e100 (2020).
424 43. Quast, C. *et al.* The SILVA ribosomal RNA gene database project: improved data processing
425 and web-based tools. *Nucleic Acids Research* **41**, D590–D596 (2013).
426 44. Newton, R. J., Jones, S. E., Eiler, A., McMahon, K. D. & Bertilsson, S. A Guide to the Natural
427 History of Freshwater Lake Bacteria. *Microbiology and Molecular Biology Reviews* **75**, 14–
428 49 (2011).
429 45. Rohwer, R. R., Hamilton, J. J., Newton, R. J. & McMahon, K. D. TaxAss: Leveraging a Custom
430 Freshwater Database Achieves Fine-Scale Taxonomic Resolution. *mSphere* **3**, e00327-18
431 (2018).
432 46. McMurdie, P. J. & Holmes, S. Waste Not, Want Not: Why Rarefying Microbiome Data Is
433 Inadmissible. *PLOS Computational Biology* **10**, e1003531 (2014).
434 47. Yeh, Y.-C., Needham, D. M., Sieradzki, E. T. & Fuhrman, J. A. Taxon Disappearance from
435 Microbiome Analysis Reinforces the Value of Mock Communities as a Standard in Every
436 Sequencing Run. *mSystems* **3**, 9 (2018).
437 48. Fisher, M. M. & Triplett, E. W. Automated Approach for Ribosomal Intergenic Spacer
438 Analysis of Microbial Diversity and Its Application to Freshwater Bacterial Communities.
439 *Appl. Environ. Microbiol.* **65**, 4630–4636 (1999).
440 49. Yannarell, A. C., Kent, A. D., Lauster, G. H., Kratz, T. K. & Triplett, E. W. Temporal Patterns
441 in Bacterial Communities in Three Temperate Lakes of Different Trophic Status. *Microb*
442 *Ecol* **46**, 391–405 (2003).
443 50. Shade, A. *et al.* Interannual dynamics and phenology of bacterial communities in a
444 eutrophic lake. *Limnol. Oceanogr.* **52**, 487–494 (2007).

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

52. Hall, M. W., Rohwer, R. R., Perrie, J., McMahon, K. D. & Beiko, R. G. Ananke: temporal clustering reveals ecological dynamics of microbial communities. *PeerJ* **5**, e3812 (2017).

53. Herren, C. M. & McMahon, K. D. Cohesion: a method for quantifying the connectivity of microbial communities. *The ISME Journal* **11**, 2426 (2017).

Figures

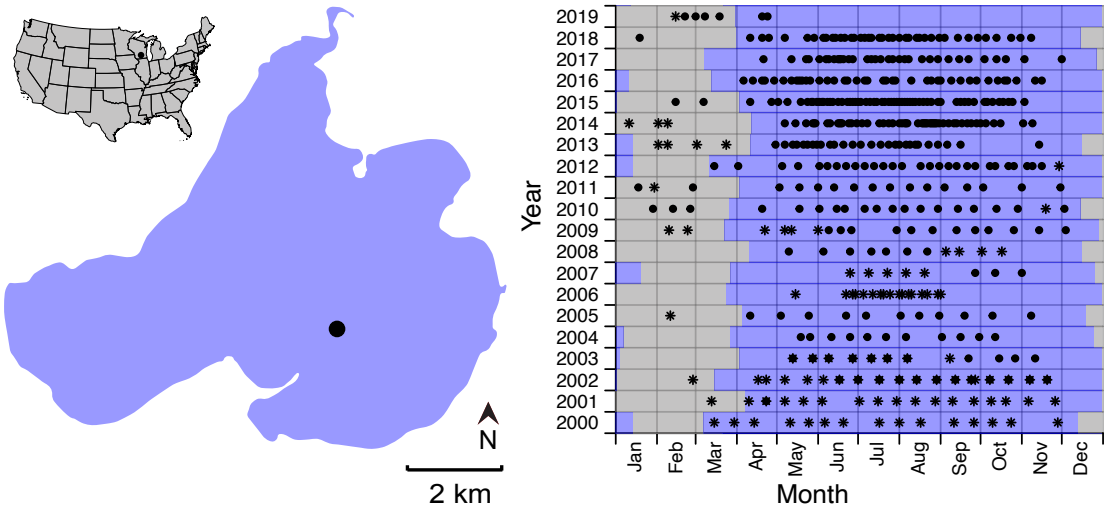

Fig 1. Sample dates and location. A: Map of Lake Mendota, with the deep hole sampling location indicated with a dot. B: Sample dates are displayed as points. White shading indicates periods of ice-on and blue shading indicates periods of ice-off. Samples points indicated with a star as opposed to a solid 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.

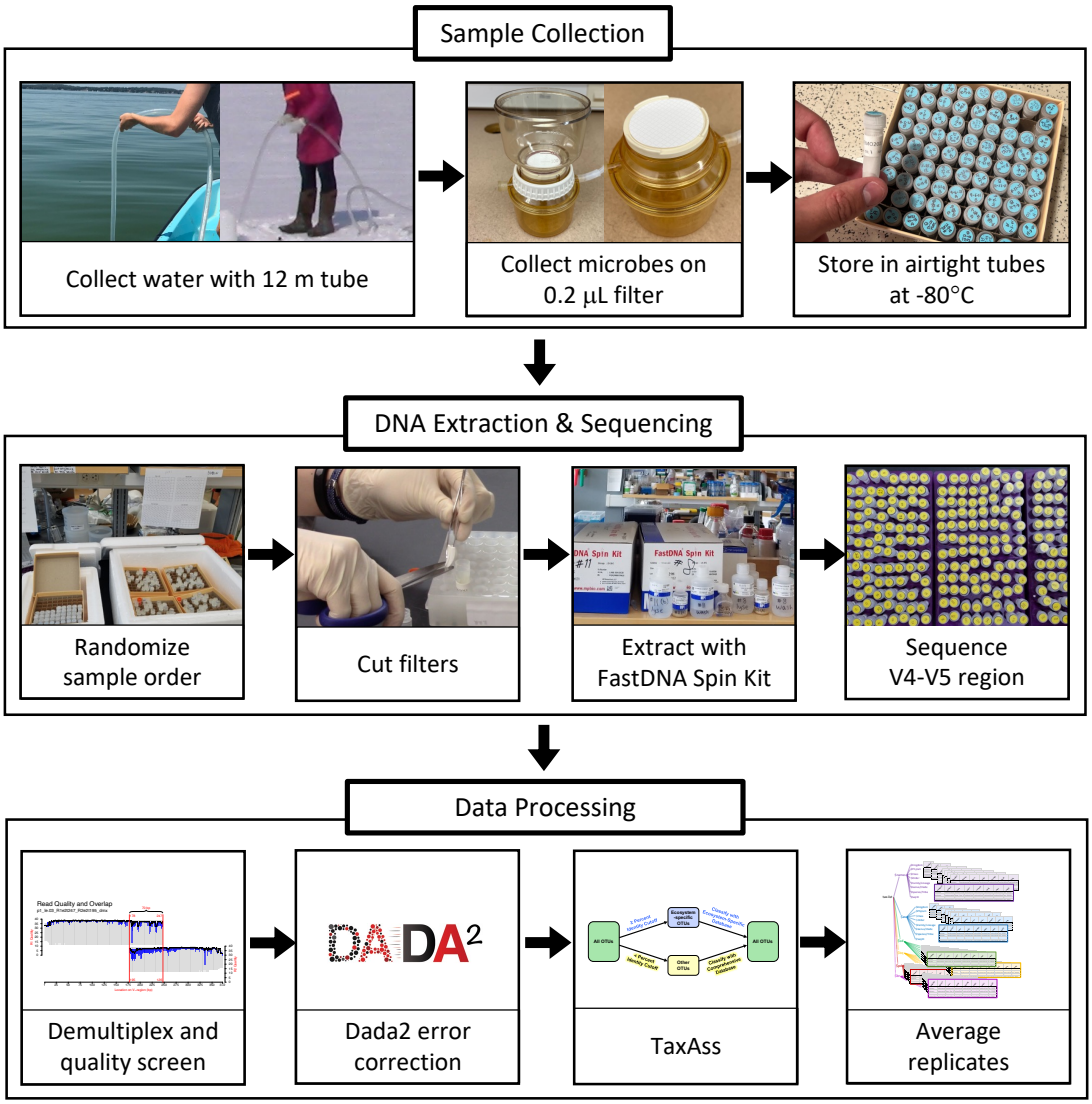

Fig 2. Methods overview. A schematic overview of the sampling methods, wet lab methods, and computational methods used to create the limony dataset.

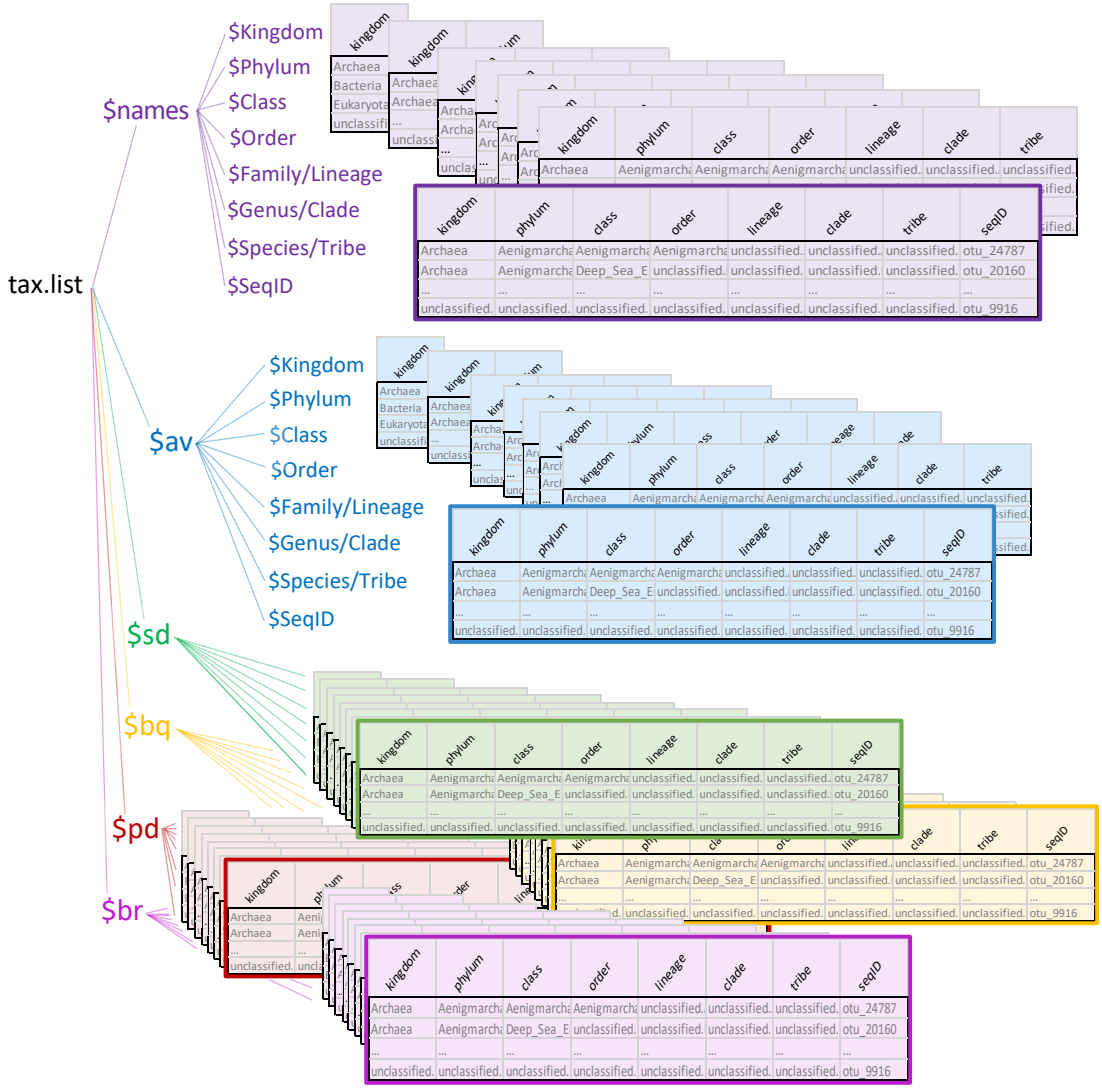

Fig 3. Limony nested list data structure. A visual representation of the nested list data structure in which processed data are provided as part of the limony R package. Abbreviations: av = average, sd = standard deviation, bq = below quantification, pd = partially detected, br = below resolution.

Tables

Online-only Table 1. Sample metadata. This table links sample names with sequencing codes, sample dates, and notes about any non-standard sample replicates, locations, depths, and filtration.