# Article information

## Article title

Environmental DNA sequencing dataset from Lake Erie algal blooms using Oxford Nanopore MinION

## Authors

Alexander F. Koeppel1, Will Goodrum2, Morgan Steffen3, Louie Wurch3, Stephen D. Turner1\*

## Affiliations

1. Signature Science LLC: Charlottesville Office, 1670 Discovery Drive, Suite 240, Charlottesville, VA 22911
2. General Atomics - Commonwealth Computer Research, Inc. (GA-CCRi): 1440 Sachem Pl., Charlottesville, VA 22901
3. James Madison University: Bioscience Building, MSC 7801, 951 Carrier Drive, Room 2001, Harrisonburg, Virginia 22807

## Corresponding author’s email address and Twitter handle

* Email: [sturner@signaturescience.com](mailto:sturner@signaturescience.com)
* Twitter: @strnr

## Keywords

* MinION
* Nanopore
* eDNA
* Harmful algal blooms
* Freshwater ecology

## Abstract

Here we describe a publicly available environmental DNA (eDNA) sequence dataset, consisting of samples collected from a National Oceanic and Atmospheric Administration (NOAA) Great Lakes Environmental Research Laboratory (GLERL) on Lake Erie. Samples were drawn before, during, and after a 2019 *Microcystis* bloom, and sequenced using 3rd generation sequencing with the Oxford Nanopore MinION device. eDNA sequences were classified taxonomically, and abundances of harmful algal bloom (HAB) -associated taxa were estimated. While the taxonomic data showed evidence of significant human and *E. coli* contamination, we found abundant *Mycrocystis*, especially in the samples drawn from bloom environments. The raw sequence data are available in the Sequence Read Archive (SRA) under accession number PRJNA812770. HABs pose a significant and increasing risk, both to human health and to the Blue Economy, and genomic approaches to early detection promise to help mitigate these risks. As such, this dataset could be of interest to freshwater ecology research teams, or any stakeholders interested in the detection and mitigation of HABs.

# Specifications table

|  |  |
| --- | --- |
| Subject | Environmental Genomics and Metagenomics |
| Specific subject area | Metagenomic analysis of freshwater microbes associated with harmful algal blooms |
| Type of data | Raw Metagenomic Data; Metadata Table; Figure |
| How the data were acquired | Freshwater eDNA samples collected from Lake Erie GLERL |
| Data format | Raw data (fastq.gz.file) |
| Description of data collection | DNeasy PowerWater® isolation kit was used to extract DNA from the water samples. DNA was then sequenced using the Oxford Nanopore MinION® device. |
| Data source location | Lake Erie NOAA GLERL stations WE13 and WE02. Latitude (approx.) 42.07 N Longitude (approx.) 81.34 W. |
| Data accessibility | The raw sequences have been deposited in the NCBI Sequence Read Archive (SRA) under accession number PRJNA812770; Data and R code for the figures and tables in this manuscript can be located at <https://github.com/signaturescience/habssed-data-ms> |

# Value of the data

* The data provide a snapshot of the microbiota present before and during a *Microcystis* bloom.
* Harmful algal blooms, including *Microcystis* cause significant harm to human health and local economies.
* Understanding the taxonomic makeup of the freshwater microbiome can give us insights into the dynamics of bloom formation.
* Any stakeholders interested in the dynamics of HABs in general, and *Microcystis* blooms specifically may find these data to be of value.

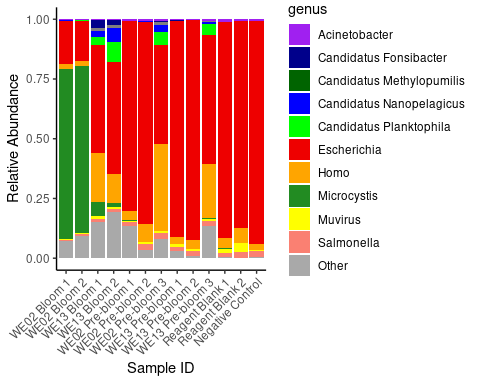
# Data Description

This dataset consists of raw eDNA reads from before and during a *Microcystis* bloom. The DNA was sequenced using the ONT MinION, and reads were classified taxonomically using the ONT What’s In My Pot (WIMP) pipeline. The complete dataset (all reads passing MinION QC) numbered 1607129 reads. Of these, 975123 were successfully taxonomically classified. Table 1 provides a subset of the metadata associated with these samples, including the GLERL from which the sample was drawn, the date of sampling, bloom condition at the time of sampling, and DNA concentration. Because some samples were sequenced as a single run, and others as part of a multiplexed run, the multiplexing information is also included. The full metadata is available on the SRA bioproject page.

Subset of the available metadata about samples sequenced in this study, available at SRA accession number PRJNA812770.

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| Sample | Run date | GLERL | Condition | Input (ng/uL) | Run type | Reads (passing QC) |
| MP1\_WE02\_B1 | 2022-01-26 | WE02 | Bloom | 22.9 | Multiplexed | 165800 |
| MP1\_WE13\_B1 | 2022-01-26 | WE13 | Bloom | 12.4 | Multiplexed | 131684 |
| MP2\_WE02\_B2 | 2022-02-16 | WE02 | Bloom | 6.16 | Multiplexed | 133788 |
| MP2\_WE13\_B2 | 2022-02-16 | WE13 | Bloom | 10.8 | Multiplexed | 164230 |
| MP2\_NC1 | 2022-02-16 | NA | Negative control | LOW | Multiplexed | 10156 |
| FR1\_WE13\_PB6 | 2021-10-07 | WE13 | Pre-bloom | 7.98 | Single sample | 609113 |
| FR2\_WE02\_PB2 | 2021-11-08 | WE02 | Pre-bloom | 0.353 | Single sample | 332514 |
| MP1\_WE02\_PB1 | 2022-01-26 | WE02 | Pre-bloom | 0.264 | Multiplexed | 10051 |
| MP1\_WE13\_PB1 | 2022-01-26 | WE13 | Pre-bloom | 0.053 | Multiplexed | 13004 |
| MP2\_WE02\_PB2 | 2022-02-16 | WE02 | Pre-bloom | 0.159 | Multiplexed | 12447 |
| MP2\_WE13\_PB2 | 2022-02-16 | WE13 | Pre-bloom | LOW | Multiplexed | 12221 |
| MP1\_RB1 | 2022-01-26 | NA | Reagent blank | LOW | Multiplexed | 5568 |
| MP2\_RB1 | 2022-02-16 | NA | Reagent blank | LOW | Multiplexed | 6553 |

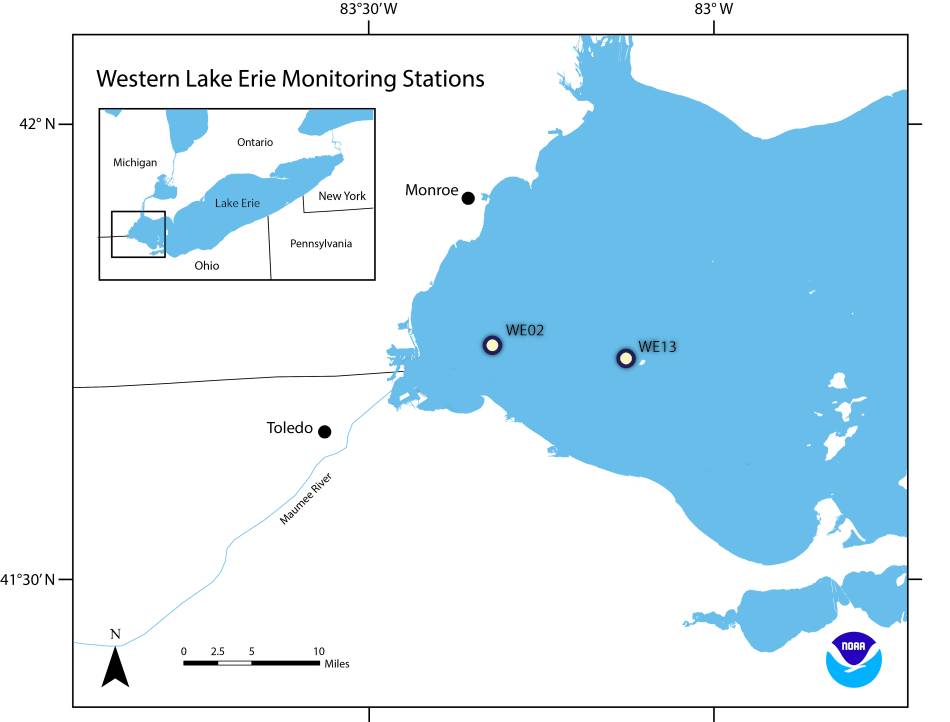
The results of the taxonomic classification are presented in figure . There was notable human and *E. coli* contamination present in several samples that was not revealed until taxonomic classification. While we took steps to track down and eliminate the contamination for later samples, we were unable to conclusively discover the source. Based on the quantities observed in the negative controls and reagent blanks, these included at minimum, 1,400 to several thousand reads classifying as *E. coli*, and from 50 to several hundred reads classifying as *H. sapiens*, and hundreds more classifying to other taxa, including *Shigella*, *Acinetobacter*, and *Microcystis* (though never more than 10 reads in any of the blanks or negative control). In some of the bloom-drawn samples (those from GLERL WE13) the abundance of *E. coli* was greater than that of *Microcystis* despite the fact that these were *Microcystis* blooms. These contaminants could therefore have had a skewing effect on the abundance results.



Stacked bar charts of relative taxon abundance (genus level) for all samples in the dataset. Top 10 most abundant genera across all samples are displayed. All other genera were classed as ‘Other.’ Sample names are the GLERL from which the sample was drawn, the condition (Bloom or Pre-bloom) and the replicate number. Reagent blank and negative control results are also inlcuded (right 3 columns). Note the presence of and classified reads (putative contamination) across all samples.

# Experimental design, materials and methods

We sequenced water samples collected from NOAA GLERL stations WE02 and WE13 during a 2019 *Mycrocystis* bloom in Lake Erie. Sampling excursions from these NOAA GLERL stations in the 2019 bloom season took place during pre-, peak, and post-bloom conditions. Samples were drawn using standard collection protocols (Steffen et al. 2017) at a water depth of 0.5m-1.5m. Metadata on the physical and chemical properties of the water at the time of sampling were also collected (Boedecker et al. 2020). Collection sites are show in the map in Figure (adapted from www.glerl.noaa.gov).



GLERL sites WE02 and WE13 from which water samples sequenced here were collected (adapted from <https://www.glerl.noaa.gov/>).

Biomass was collected onto 0.2 um Sterivex filters and kept on ice until they were returned to the lab where they were stored at -80C until extraction. Extraction protocol was adapted according to (Cruaud et al. 2017). We then prepared the libraries for MinION sequencing using Oxford Nanopore Technologies (ONT) sequencing kit (initially the Rapid Sequencing Kit (SQK-RAD004) but transitioning to the Rapid Barcoding Kit 96 (SQK-RBK110.96) for later runs). DNA cocentrations for each sample were estimated using a Qbit analyzer. We sequenced the eDNA from those samples using the ONT MinION Mk1C device, which performed base-calling and quality filtering using ONT’s embedded MinKNOW software using the default settings. We then performed taxonomic classification using the ONT What’s In My Pot (WIMP) pipeline. Relative abundance was computed by dividing these genus-specific read counts by the total number of reads passing the QC filters for the same samples.

Figure was generated in R(R 2021), by joining the WIMP taxonomic results for all samples to their respective NCBI taxonomic IDs, aggregating at the genus level, and then plotting. The code used to generate the figure can be found at: <https://github.com/signaturescience/habssed-data-ms>

# Ethics Statements

The authors declare no financial or commercial conflicts of interest.

# CRediT author statement

**Alex Koeppel**: Performed sequencing and conducted all data analysis. **Will Goodrum, Morgan Steffen, Louie Wurch, Stephen D. Turner**: Contributed to study design, results interpretation, and drafting and reviewing the manuscript.

# Acknowledgements

The authors would like to express sincere gratitude and acknowledge Jessica Bouchet, David Russell, and Hunter Baylous for assisting with laboratory work on this project, and Dr. Stephanie Guertin for consultation on design and execution.

This material is based upon work supported by the SBIR Program within the NOAA Technology Partnerships Office under Grant No. NA21OAR0210481.

# References

Boedecker, Ashlynn R., Desi N. Niewinski, Silvia E. Newell, Justin D. Chaffin, and Mark J. McCarthy. 2020. “Evaluating Sediments as an Ecosystem Service in Western Lake Erie via Quantification of Nutrient Cycling Pathways and Selected Gene Abundances.” *Journal of Great Lakes Research* 46 (4): 920–32. <https://doi.org/10.1016/j.jglr.2020.04.010>.

Cruaud, Perrine, Adrien Vigneron, Marie-Stéphanie Fradette, Steve J. Charette, Manuel J. Rodriguez, Caetano C. Dorea, and Alexander I. Culley. 2017. “Open the Sterivex Casing: An Easy and Effective Way to Improve DNA Extraction Yields.” *Limnology and Oceanography: Methods* 15 (12): 1015–20. <https://doi.org/10.1002/lom3.10221>.

R, Core Team. 2021. “R: A Language and Environment for Statistical Computing.” *R Foundation for Statistical Computing* Vienna, Austria. <https://www.R-project.org/>.

Steffen, Morgan M., Timothy W. Davis, R. Michael L. McKay, George S. Bullerjahn, Lauren E. Krausfeldt, Joshua M. A. Stough, Michelle L. Neitzey, et al. 2017. “Ecophysiological Examination of the Lake Erie Microcystis Bloom in 2014: Linkages Between Biology and the Water Supply Shutdown of Toledo, OH.” *Environmental Science & Technology* 51 (12): 6745–55. <https://doi.org/10.1021/acs.est.7b00856>.