Dataset key for manuscript “Harmful algal blooms preceded by a predictable and quantifiable shift in the oceanic microbiome”

Miranda C. Mudge1,2, Michael Riffle1, Gabriella Chebli3, Deanna Plubell1, Tatiana Rynearson4, William S. Noble1, Emma Timmins-Schiffman1, Julia Kubanek3, Brook L. Nunn1\*

1Department of Genome Sciences, University of Washington, Seattle, WA

2Department of Molecular and Cellular Biology, University of Washington, Seattle, WA

3School of Biological Sciences and School of Chemistry & Biochemistry, Georgia Institute of Technology, Parker H. Petit Institute for Bioengineering and Bioscience, Atlanta, Georgia

4Graduate School of Oceanography, University of Rhode Island, Kingston, RI

**Dataset Descriptions**

**Dataset 1: Environmental Metadata**

Water chemistry measurements during full field season (May 27 – June 18, 2021). Dataset contains metadata matched to the time of collection of bacterial fraction of the water column (every 4 hours). Columns include: Date, Time, DateTime, Cyanobacteria, Picoeuks, Nanoeuks, Intensity….lum.ft2, Chlorophyll.RFU, Conductivity.uscm, ODO.sat, Sal.psu, pH, Temp.C, PO4, Si.OH.4, NO3, NO2, and NH4. Cyanobacteria, Picoeuks (picoeukaryotes), and Nanoeuks (nanoeukaryotes) represent flow cytometry data collected on various microbial size fractions reported in cells/mL. Intensity represents light levels collected by HOBO logger device deployed at site of sample collection. Chlorophyll.RFU, Conductivity.uscm, ODO (dissolved oxygen, % saturation), Sal.psu (salinity), pH, and Temp.C (temperature degrees Celsius) were measures taken by YSI EXO1 Sonde probe deployed at collection site. PO4 (phosphate), Si.OH.4 (silicate), NO3 (nitrate), NO2 (nitrite), and NH4 (ammonia) are the results (in mM) of total nutrient analysis.

**Dataset 2: DDA-refined database**

FASTA file of all proteins from the data-dependent acquisition (DDA)-refined database (DB). This file was used to create a chromatogram library for searching files in the data-independent acquisition (DIA) experiment.

**Dataset 3: Peptide abundance time series**

Peptide abundances for all peptides identified in the data-independent acquisition (DIA) experiment across the discovery time series from June 12 (17:00) – June 18 (13:00). Column 1 contains peptide identifications. The columns represent samples, denoted with day and hour of collection separated with a period (*e.g.,* 16.17 represents the sample collected on day 16 at hour 17:00).

**Dataset 4:** **Phase 1A peptides - PC2 loadings**

PC2 loadings for peptides identified by Phase 1A. The top 10% of peptides (447 peptides) driving variation across PC2 in the PCA analysis of Phase 1A chosen to proceed to Phase 2. PC loading reported along with peptide identification.

**Dataset 5:** **MetaGOmics peptide abundance time series**

MetaGOmics analysis identified class-level taxonomic assignment for all peptides with known annotation (2015 total peptides). Class denoted by column “taxonomy\_class”. Time series of peptide abundances across discovery dataset included (refer to Dataset 3 for all peptides).

**Dataset 6:** **Phase 2 biomarker peptides and HMM state shift timing**

All 314 biomarker candidate peptides identified by Phase 2. Peptide identification and time point (day.hour) immediately post-state shift are reported. Column “Time\_HMM\_SS) reported by a simple two state hidden Markov model (HMM) fit to time series of peptide abundance for each individual peptide identified in Phase 1 of the discovery analysis.

**Dataset 7:** **Spearman's correlation for biomarker peptides between blooms**

Spearman’s correlation coefficient for each peptide abundance across the discovery and validation time series. The bloom start date (BSD) was calculated as the point of greatest change in chlorophyll a concentration prior to the bloom peak. A hidden Markov model was used to calculate the precise timing of the underlying state shift in peptide abundance. Peptides abundances were aligned by timing of their respective BSD, working backwards. A total of 24 time points were considered to keep samples equivalent. A Spearman’s correlation coefficient rho value and corresponding p-value were calculated for the peptide abundances between the discovery and validation time series.

**Dataset 8: Biomarker FASTA**

Word file of FASTA entries for the final 12 peptides validated as HAB onset biomarkers. FASTA entries are proteins, and individual peptide sequences are highlighted.

**Dataset 9: Verrucomicrobiae estimated gene count time series**

Estimated gene copies for the bacterial class Verrucomicrobiae across the full time series. “TimepointDate” and “Estimated\_gene\_count\_Verrucomicrobiae” are reported. Estimated gene copies are gene counts identified in whole shotgun metagenomic sequencing from the collected microbiome that provide rough estimates of taxonomic groups based on specificity of the 150BP sequence.