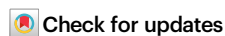


Harmful algal blooms are preceded by a predictable and quantifiable shift in the oceanic microbiome

Received: 7 June 2024

Accepted: 16 April 2025

Published online: 29 April 2025



Miranda C. Mudge¹, Michael Riffle¹, Gabriella Chebli², Deanna L. Plubell¹, Tatiana A. Rynearson³, William S. Noble¹, Emma Timmins-Schiffman¹, Julia Kubanek^{2,4} & Brook L. Nunn¹✉

Harmful algal blooms (HABs) have become a worldwide environmental and human health problem, stressing the urgent need for a reliable forecasting tool. Dynamic interactions between algae, including harmful algae, and bacteria play a large role regulating water chemistry. Free-living bacteria quickly respond to small physical and/or chemical environmental changes by adjusting their proteome. We hypothesize that this response is detectable at the peptide level and occurs before rapid phytoplankton growth characteristic of harmful bloom events. To characterize the microbiome's physiological changes preceding bloom onset, we collected and analyzed a high-resolution metaproteomic time series of a free-living microbiome in a coastal ecosystem. We confirm that twelve candidate HAB biomarkers are detectable, quantifiable, and correlated across two pre-bloom periods. This study identifies proteomic shifts in bacterial peptides which may be used as predictive biomarkers for forecasting harmful algal bloom initiation, potentially mitigating detrimental algal bloom outcomes in the future.

Harmful algal blooms (HABs) are sudden accumulations of photosynthetic phytoplankton that cause adverse effects to aquatic ecosystems¹. These blooms pose serious threats to both human health and the economy² with negative consequences on fishing, tourism, and real estate industries as well as mass mortalities of coastal marine mammals, seabirds, and fishes. Harmful effects caused by bloom-forming species are primarily attributed to the production of highly toxic compounds such as neurotoxins³; that said, there are many non-toxic HAB species that achieve remarkably high cell densities that can block sunlight to deeper waters, alter local water chemistry (e.g., pH, nutrients)⁴, generate hypoxic or anoxic zones⁵, and/or obstruct fish gills, all potentially leading to a swift rise in marine animal mortalities⁶. A variety of human-made and environmental perturbations are capable of initiating these blooms⁷; however, in the majority of bloom

events, the specific underlying cause for HAB initiation is unknown. In general, two fundamental goals of HAB research are (1) to determine what factors lead to the dominance of HAB-forming species and (2) to predict bloom onset, enabling warning and mitigation strategies to combat the increasing frequency and intensity of HAB events globally⁸.

Interactions between phytoplankton and bacteria have been shown to play a paramount role in shaping the dynamics of nutrient availability, biogeochemical cycles, and microbial diversity and function in marine ecosystems^{9–11}. The individual and interacting factors influencing HAB initiation, including chemical, physical, and biological aspects of an ecosystem, are experienced by both HAB-forming phytoplankton as well as their bacterial neighbors. In the specific context of HABs, tight associations have been observed within the microbiome as phytoplankton advance through bloom cycles, with bacteria playing

¹Department of Genome Sciences, University of Washington, Seattle, WA, USA. ²School of Biological Sciences, Georgia Institute of Technology, Parker H. Petit Institute for Bioengineering and Bioscience, Atlanta, GA, USA. ³Graduate School of Oceanography, University of Rhode Island, Kingston, RI, USA. ⁴School of Chemistry & Biochemistry, Georgia Institute of Technology, Parker H. Petit Institute for Bioengineering and Bioscience, Atlanta, GA, USA.

✉ e-mail: brookh@uw.edu

a deterministic role in the outcome of HAB progression and algal byproducts^{12–14}. Additionally, certain consistently abundant taxa have been noted as commonly associated with different phytoplankton species^{15–17}, lending precedent to the idea that community structure and function are driven by specific associations¹⁸. The influence these organisms have on each other's biological potential creates a recognized opportunity¹⁹ to utilize bacterial responses as a rapid in situ predictor for phytoplankton dynamics, regardless of the causative influence that may trigger bloom initiation.

Proteomic techniques are capable of capturing this active functionality of cells through the measurement of peptides, the component pieces of proteins that are tightly regulated and rapidly modified in response to minute environmental perturbations^{20,21}. Recent advances in mass spectrometry (MS) techniques and methods have enabled highly accurate quantification of peptides in a variety of sample types across a wide dynamic range²². Additionally, as the development of peptide-based biomarkers for diagnosing human health²³ is gaining momentum in medical research, a range of potential molecular or antigen rapid tests could be deployed via citizen science²⁴ or autonomous vehicle monitoring for environmental applications²⁵. While proteomic analysis of eukaryotic organisms like phytoplankton remain difficult²⁶, a predictive biomarker developed using a bacteria-centric approach would potentially eliminate the need for targeted forecasting due to taxonomic and functional overlap of marine microbiomes globally^{27,28}.

Here we present a high-resolution time series that captures the biological and chemical characteristics of a dynamic coastal ecosystem during two pre-bloom periods of *Chaetoceros socialis*, a cosmopolitan and abundant HAB-forming diatom capable of deteriorating water quality and killing fish through physical irritation²⁹. To interrogate the microbiome for biomarkers capable of predicting a HAB > 24 hours in advance, we perform a metaproteomic analysis that quantifies bacterial peptides collected every four hours. A multistep framework is developed for selecting and statistically validating potential biomarker peptides using a hidden Markov model (HMM), a statistical approach that can identify non-reverting state shifts in time series data. Our investigation reveals 314 potential biomarkers for predicting bloom onset. Targeted validation of peptides from a second pre-bloom period with a different chemical profile confirms a subset of 12 peptides as quantifiable, non-reverting biomarkers. These findings provide evidence towards the use of the bacterial microbiome to generate biomarkers for predicting future HABs.

Results

Field-sampling captures two pre-bloom initiation periods with divergent nutrient profiles

A high-resolution time series of a coastal bacterial microbiome was acquired from May 27 to June 18, 2021, in East Sound, WA, USA, capturing samples every 4 hours across one full HAB cycle and the pre-bloom dynamics of a second HAB (Fig. 1a; Supplementary Fig. 1; Supplementary Data 1). On-site microscopy determined that the eukaryotic fraction of the ecosystem was dominated by *Chaetoceros socialis* (SoundToxins Phytoplankton Monitoring Network)³⁰. Two bloom initiations of *C. socialis* were identified from the detailed chlorophyll fluorescence trace (Fig. 1b, c). Calculated bloom start dates (BSD)³¹, defined as the time point with the highest positive change in chlorophyll *a* per change in time preceding the bloom peak, were May 31 at 23:00 and June 18 at 11:00 (Supplementary Fig. 2). Cell counts of nanoeukaryotes and picoeukaryotes confirmed the May 31st BSD³⁰ (Fig. 1d), while Sound Toxin microscopy and satellite images confirmed the presence of a HAB on June 19 [SoundToxins Phytoplankton Monitoring Network]³⁰ (Fig. 2). The nutrient conditions over the two pre-bloom initiation periods revealed divergent chemistries (Fig. 1h, i). Specifically, the pre-bloom period in May experienced a 20–30 mM increase in silicate and a 15 mM increase in nitrate ~60 hours in advance

of the BSD (Fig. 1h), coinciding with a 4 °C fluctuation in temperature (Fig. 1j). In contrast, in the pre-bloom conditions prior to the June BSD concentrations of all nutrients measured fluctuated <5 mM, revealing a more stable chemical ecosystem (Fig. 1i).

As a result of these divergent nutrient profiles in the pre-bloom conditions, the two time series were considered as separate in situ pre-bloom experiments (discovery and validation), allowing the opportunity to characterize each ecosystem independently (Fig. 1a). June 12 to 18 was selected as the discovery period because the environmental parameters monitored were more stable (Fig. 1f). May 28 to June 2 was selected as the validation period to test candidate biomarkers (Fig. 1a) to confirm that peptides remained non-reverting at least 24 hours post-BSD.

Discovery phase: identifying peptide biomarkers through quantitative metaproteomics

A metaproteomic pipeline was utilized to 1) confidently identify peptides from this complex microbiome, 2) generate quantitative peptide data, and 3) annotate peptides using a lowest common ancestor approach.

Briefly, data-dependent acquisition (DDA) MS data from the June 12 to 18 sample set (36 samples) were searched against the metagenome-generated database (DB) of 3,457,818 proteins (PRIDE identifier PXD052873). DDA analysis yielded 36,429 protein detections from the 9765 peptides confidently detected, which were used to generate the smaller DDA-refined database for quantitative data independent acquisition (DIA) analysis and interpretation (Supplementary Fig. 3; Supplementary Data 2)^{32,33}. The resulting DDA-refined database represents 1% of the original metagenome. Next, the DIA analysis of the discovery sample set detected 4645 unique peptides. A conservative approach was taken to only include peptides that yielded a quantitative value for > 50% of the 36 time points, yielding a final set of 4468 peptides for the next phase of our biomarker screen (Fig. 3; Supplementary Data 3).

For Phase 1A, a principal component analysis (PCA) was performed on the time series to identify how the quantified peptides varied by time and day of collection. A PCA of all quantified peptides from the DIA analysis of the discovery period revealed that variation across PC1 (proportion of variance = 12.67%) was driven by absence of peptide signals across the time series (Supplementary Fig. 4). Visualization of PC2 (proportion of variance = 9.17%) vs PC3 (proportion of variance = 6.67%) revealed two clusters of data separating along PC2 (Fig. 4a). Analysis of the date and time of collection represented by each datapoint revealed that the separation was driven by the date of collection (between June 15–16; Fig. 4a), not the time of day of collection as expected. A Mann-Whitney test of PC2 loadings between June 12 to 15 and June 16 to 18 was significant ($p = 1.916 \times 10^{-9}$). A total of 447 peptides, representing the top 10% of peptides driving variation across PC2, were selected for Phase 2 analysis (Supplementary Data 4).

In Phase 1B, 2015 of the total 4468 peptides were annotated with MetaGOmics³⁴, revealing representation of 29 total bacterial classes in the quantitative peptide dataset (Supplementary Fig. 5; Supplementary Data 5). Nearly half of those annotated peptides were from Alphaproteobacteria (956 peptides). The next six most abundant classes were Flavobacteria (431 peptides), Gammaproteobacteria (339 peptides), Verrucomicrobiae (91 peptides), Bacilli (75 peptides), Opitutae (39 peptides), and Betaproteobacteria (14 peptides) (Supplementary Table 1). Seven of the 29 taxa showed significant correlation between summed quantitative peptide abundance values and time over the pre-bloom period (Spearman's rho cut-off = |0.5|; Fig. 4c; Supplementary Table 2). Verrucomicrobiae, Opitutae, Acidobacteriia (1 peptide), and Zetaproteobacteria (1 peptide) exhibited a strong positive trend over time ($\rho \geq 0.7$, Fig. 4c), with a distinct inflection in abundance at the same June 15–16 changepoint identified from the PCA (Fig. 4a, c). The peptide trend from the most abundant class, Alphaproteobacteria, exhibited a weak positive correlation

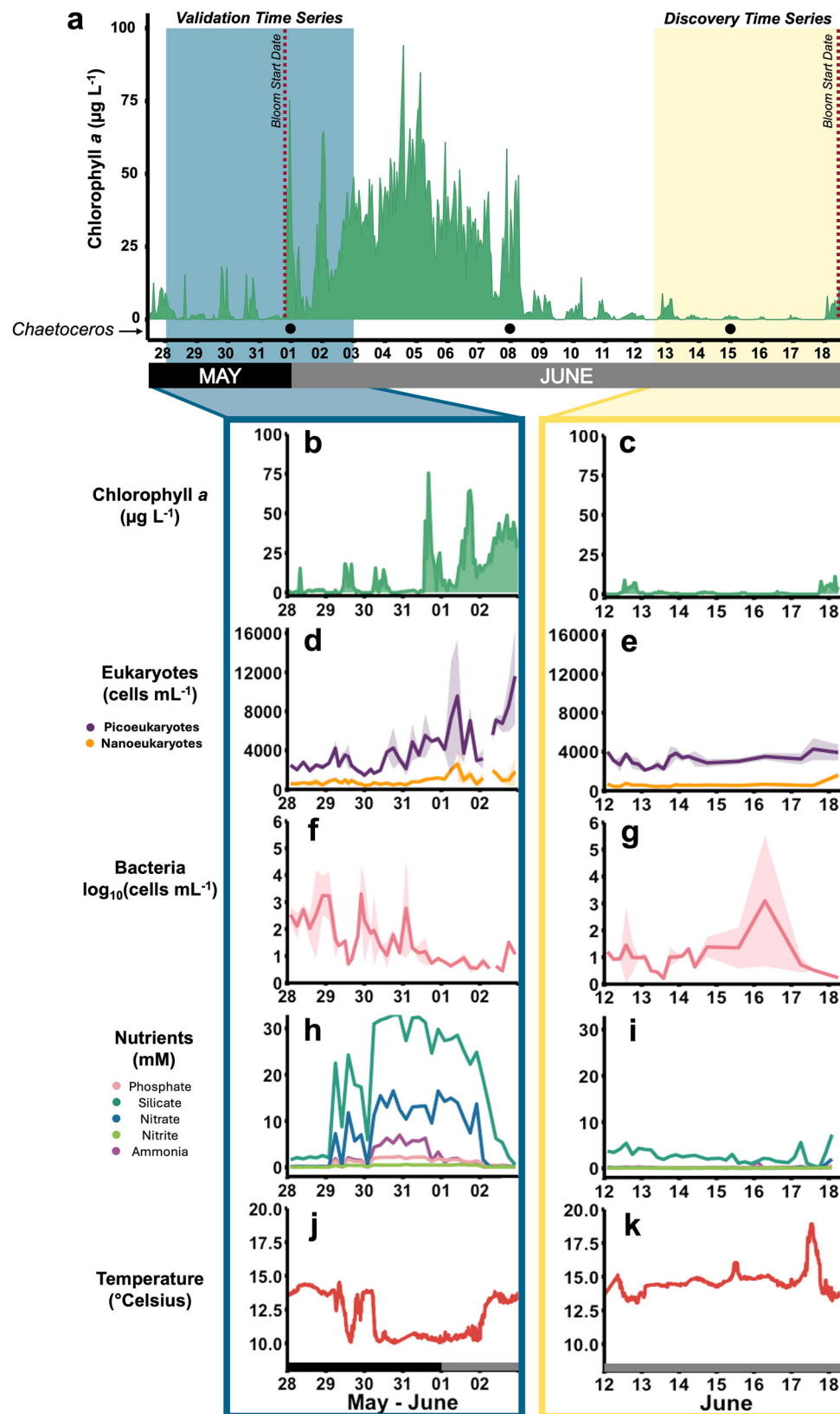


Fig. 1 | Fieldwork captures pre-bloom dynamics of two *Chaetoceros socialis* blooms. **a** Relative phytoplankton levels were captured by chlorophyll *a* measured every 10 min via YSI Exo1 Sonde probe. Time periods selected for discovery (yellow) and validation (blue) are highlighted and bloom start date (BSD), indicated with a dotted red line. Black dots denote time points where SoundToxins confirmed *Chaetoceros* spp. was the dominant phytoplankton in the water. SoundToxins also confirmed *Chaetoceros* spp. dominance on May 25, 2021, and June 22, 2021. **b, c** The

chlorophyll *a* trace for the discovery and validation datasets are expanded for alignment. **d, e** Flow cytometry of nanoeukaryotes and picoeukaryotes (in biological triplicate), as well as bacterial (in biological duplicate), **f, g** cell counts were determined using cell size every four hours. Data are presented as mean values with $\pm$ SD visualized with shading. **h, i** Nutrient concentrations were reported from total nutrient analysis every four hours. Temperature **j, k** of the water column was captured with a YSI probe. Source data are provided as a Source Data file.

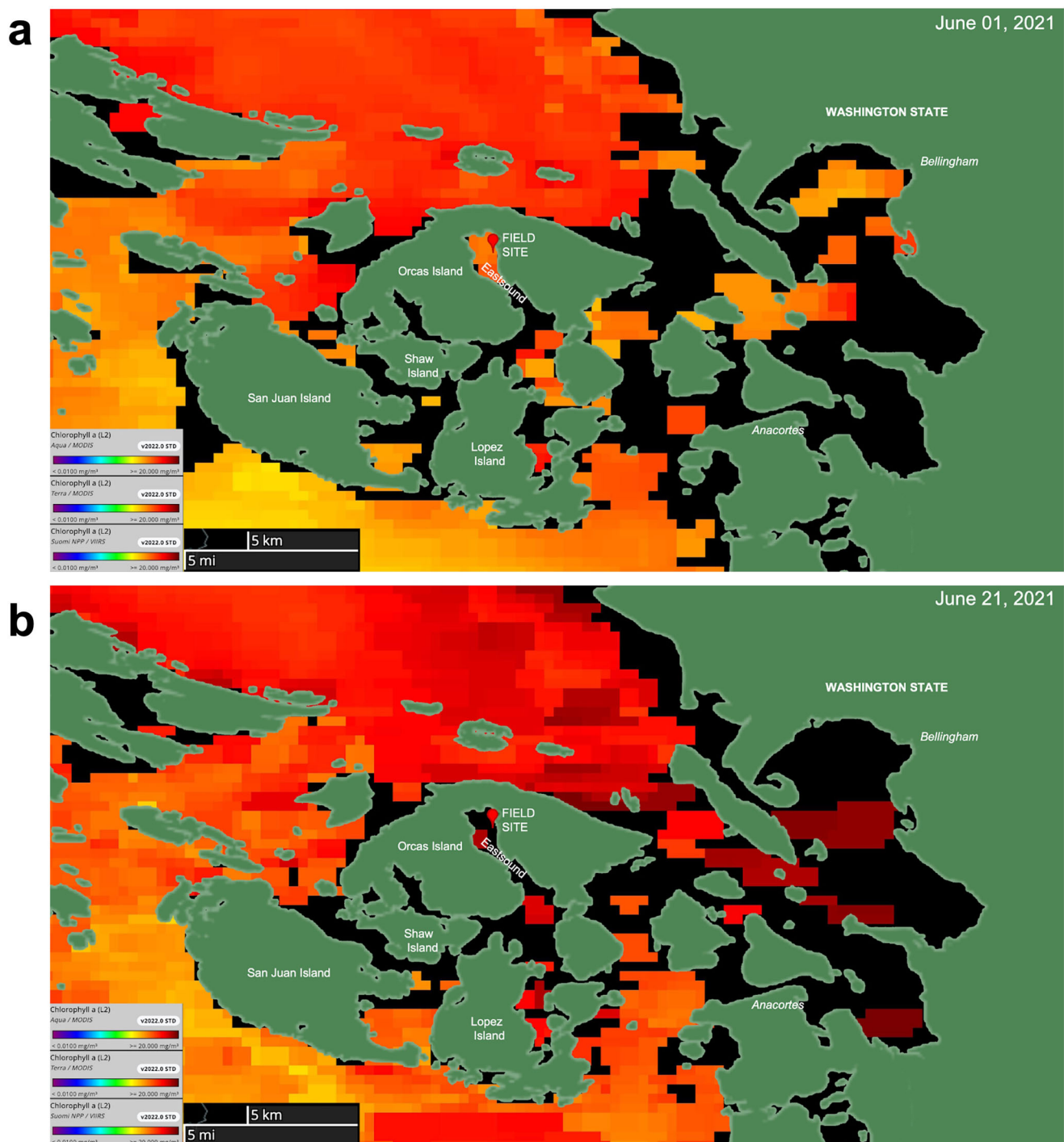


Fig. 2 | Satellite images of chlorophyll *a* concentration confirming HAB presence. MODIS satellite images of near-surface chlorophyll *a* concentration (mg m⁻³) in waters surrounding East Sound, WA, USA during (a) June 1 and (b) June 21, 2021. The MODIS Chlorophyll *a* product is available from both the Terra and Aqua satellites. The sensor and imagery resolution is 1 km, and the temporal resolution is

daily. All chlorophyll *a* layers were visualized at the specific date of interest and relative locations annotated using satellite imagery from reference: <https://go.nasa.gov/4kfUJHdd>. We acknowledge the use of imagery from the NASA Worldview application (<https://worldview.earthdata.nasa.gov>), part of the NASA Earth Science Data and Information System (ESDIS).

($\rho = 0.35$, Fig. 4c), while peptide trends from Gammaproteobacteria, Flavobacteria, and Epsilonproteobacteria (6 peptides) exhibited a negative correlation with time ($\rho \leq -0.5$, Fig. 4c). All peptides from classes with a $\rho \geq |0.5|$ were selected for the next phase of biomarker discovery. Additionally, peptides from the most abundant class, Alphaproteobacteria, were also selected to proceed through biomarker development. In total, our taxa-driven discovery approach yielded 1858 peptides for consideration as biomarkers in Phase 2.

As our goal was to identify biomarkers to predict HAB initiation, a hidden Markov model (HMM) approach was applied to reveal peptides with a non-reverting shift in underlying state processes. A two-state HMM was fit to each of the 2,015 Phase 1 candidate peptides' temporal abundances (Fig. 3, yellow). A total of 314 peptides were confirmed to undergo a non-reverting state shift >24 hours prior to the bloom start date (Fig. 4b; Supplementary Data 6). This set included 88 peptides that were identified from both Phase 1 analyses, as well as 181 that

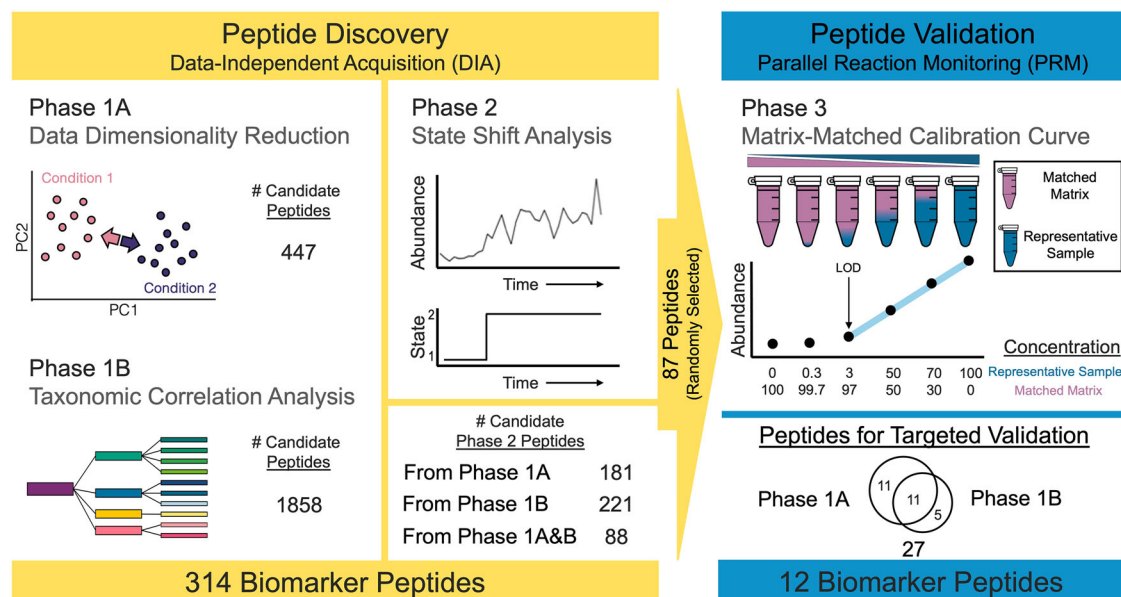


Fig. 3 | Framework for determining candidate peptides. Schematic of the biomarker identification pipeline for systematic peptide discovery (yellow-framed box) and validation (blue-framed box). For peptide discovery, data-independent acquisition (DIA) was used to quantify peptides on the mass spectrometer from the June 12 to 18 samples. Peptides contributing to pre-bloom microbiome proteome variation were identified in Phase 1A through dimensionality reduction (PC loadings indicated by pink and purple arrow illustrations in PCA plot) and Phase 1B through correlation analysis of taxonomic trends over time which identified peptides from taxonomic classes that had a time-specific correlation ($\rho \geq |0.5|$). The

candidate peptides identified in Phase 1 progressed to Phase 2, the state shift analysis using hidden Markov models. In the discovery phase, a total of 314 peptides were identified as potential biomarker candidates for downstream testing. A subset of 87 of the candidate peptides were randomly selected and to determine if they yielded quantitative linear responses in Phase 3 using a matrix-matched calibration curve. A final, randomly selected subset of 27 peptides that passed Phase 3 was evaluated using targeted analysis to assess consistency in their temporal abundance from May 28 to June 2 and to validate the biomarkers as predictors of HAB initiation.

originated from only the Phase 1A PCA and 221 from only the Phase 1B correlation analysis (Fig. 3).

From the 314 peptides that passed Phase 2, 87 peptides were randomly selected to test for a quantitative response in a dilution series using targeted mass spectrometry³⁵. Because each microbiome sample carries a complex mixture of thousands of peptides, determining the quantity of any individual peptide abundance within the sample can be difficult. In order to ensure that the selected biomarker peptides are both detectable and quantifiable when present in complex mixtures, we created serial dilutions of the sample to generate a calibration curve prior to further analysis (Fig. 2, blue). Because there are interfering ions simultaneously eluting from the chromatography column prior to MS analyses, which can influence detectability of ions and the resulting quantities reported, these serial dilutions were generated with a complex suite of peptides, rather than water. In this dilution series, the total amount of protein injected for each MS run remained the same while the ratio of microbiome sample to diluent was varied incrementally. The diluent used here consisted of a protein digest of *Colwellia psychrerythraea* 34H, a model bacterium with no shared endogenous peptides. By diluting our test microbiome with peptides of interest against a background of equivalent biochemical complexity with no shared endogenous peptides, we were able to determine if candidate biomarkers exhibited a linear response during targeted MS experiments and establish a limit of detection (LOD) for each peptide. This method, the matrix-matched calibration curve (MMCC), is employed for peptide biomarker development in complex systems³⁵. Samples prepared with the MMCC method were analyzed using parallel reaction monitoring (PRM) mass spectrometry to determine the LOD for 87 peptides down to detectable threshold of $1e3$ and to assess whether each peptide's dilution series exhibited a linear response above this threshold (Supplementary Fig. 6). Due to sample limitation, a random subset of 27 peptides that passed the MMCC standards were selected to progress through to the validation phase.

Validation of biomarkers as predictors of HABs

In order to determine whether our final subset of 27 biomarkers could predict a HAB, targeted MS analyses were performed on the May 28 - June 2 pre-bloom sample set (i.e., the validation period, Fig. 1c). The goal of the validation experiment was to evaluate whether the 27 biomarkers identified in the discovery bloom data set also exhibited a non-reverting state-shift prior to the June 18 BSD and whether the resulting temporal patterns were consistent across multiple blooms. All peptides identified as potential pre-bloom biomarkers were above the desired limit of detection (LOD) in the validation time series. A two-state HMM was fit to each peptide, revealing a total of 12 peptides that exhibited a non-reverting state shift prior to the BSD (Fig. 5a, b; Supplementary Figs. 7-8). The state shift for all peptides occurred at least 24 hours prior to the observed BSD for both blooms (Figs. 1c, 5b; Supplementary Figs. 7-8; Supplementary Table 3). Additionally, a Spearman correlation analysis of the discovery and validation blooms showed a strong positive correlation for eight of the quantified peptides when peptide abundance was aligned to the BSD (correlation value ≥ 0.6 (e.g., Fig. 5a, b; Supplementary Fig. 9; Supplementary Data 7). A detailed annotation of the peptides with Basic Local Alignment Search Tool (BLAST) protein analysis revealed that, of the final 12 validated biomarkers, 8 peptides were from Gammaproteobacteria, 2 peptides were from Flavobacteriia, and 2 peptides were from Verrucomicrobiae. Functionally, 6 of the biomarker peptides were TonB-dependent receptors, 2 peptides were ribosomal, and others are indicative of an active response to the environment (Table 1; Supplementary Data 8).

Discussion

In this study, we utilized high-temporal resolution sampling of a bacterial microbiome to reveal that certain bacterial peptides undergo explicit and reproducible state shifts ≥ 24 hours prior to the BSD of two HABs occurring in divergent chemical ecosystems. These peptides

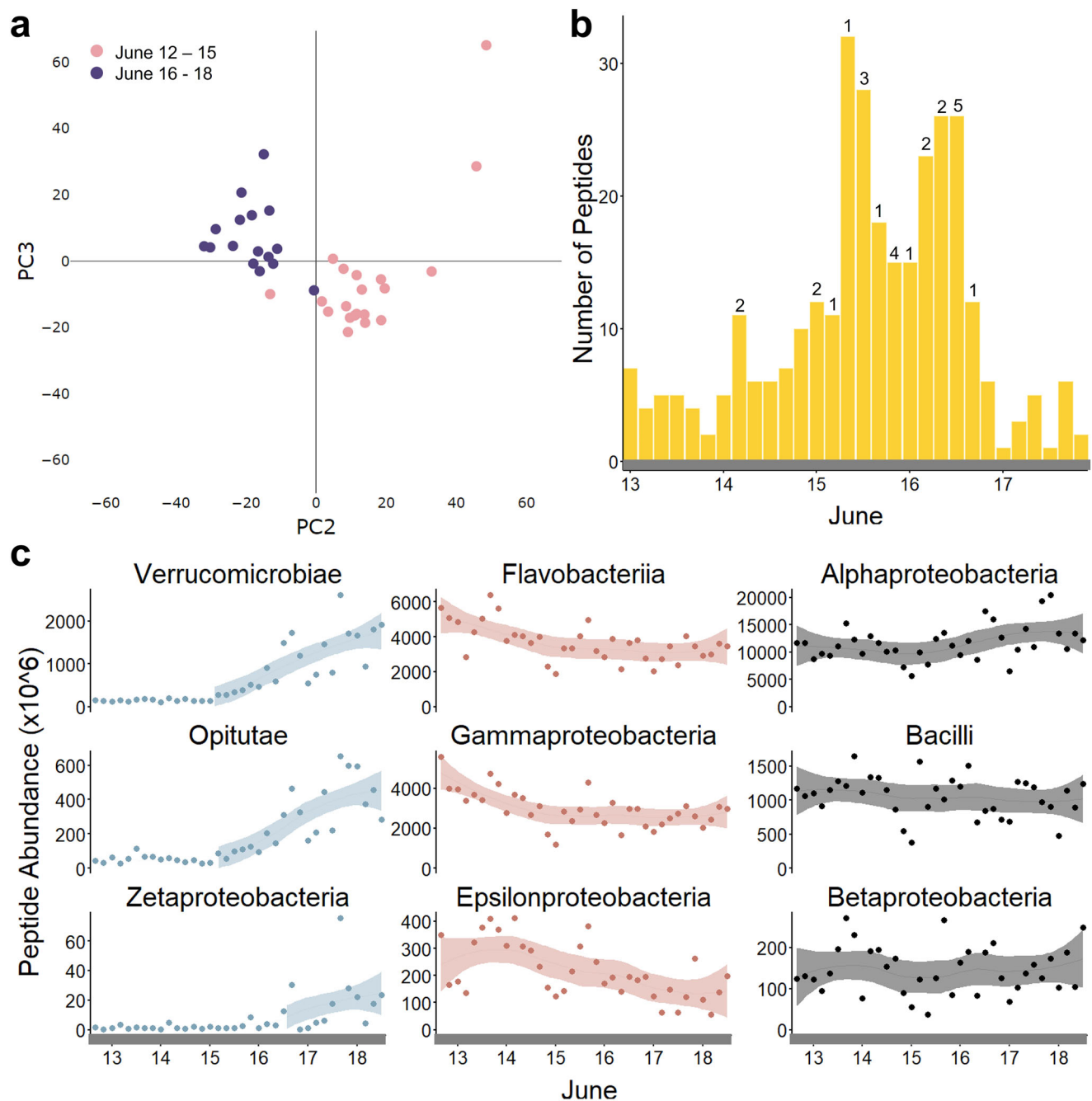


Fig. 4 | Quantified peptides characterized across the microbiome reveal trends in global structure and specific taxa. **a** PCA of all quantified bacterial peptide abundances from DIA analyses, colored to illustrate the first (June 12–15, purple) and second (June 16–18, pink) half of the time series, with PC2/3 reported. **b** Distribution of the HMM state shift timing for all non-reverting peptides when average peptide abundance before and after the state shift are compared. Numbers above the bars indicate how many peptides from that time were selected as final biomarkers for

validation. **c** Time series of the total peptide abundances of bacterial classes. Spearman's correlation coefficient ρ was calculated using each class's summed peptide abundance against ranked time values (positive correlation: ≥ 0.5 , blue, negative correlation: ≤ -0.5 , red, and no correlation: black), with shaded regions representing confidence interval using loess smoothing. Source data are provided as a Source Data file.

originated from bacterial community members whose population densities also reflected changes in total peptide abundance approximately 24 hours prior to HAB onset. These findings confirm our hypothesis that HABs and their associated microbiomes exhibit ecological coupling of metabolic activity in response to the biological or environmental factors that drive HAB initiation, allowing for the development of HAB predictive tools in the form of peptides.

As the long-term goal of this research is to identify candidate peptide biomarkers that can be easily isolated and identified in future applications, sample collection targeted free-living marine

bacteria. Targeting free-living bacteria increases the likelihood that any relevant biomarkers would be applicable for predictive use in other bloom scenarios (in comparison to particle-associated bacteria that would be specific to their respective bloom-forming phytoplankton), as well as avoids the practical complications of extracting bacterial signals from abundant phytoplankton biomass^{36,37}. Additionally, probe-based hourly chlorophyll *a* abundance was used to calculate the BSD³¹, streamlining the assessment of HAB onset independent of laborious, and sometimes unreliable, methods used to quantify cell counts³⁸.

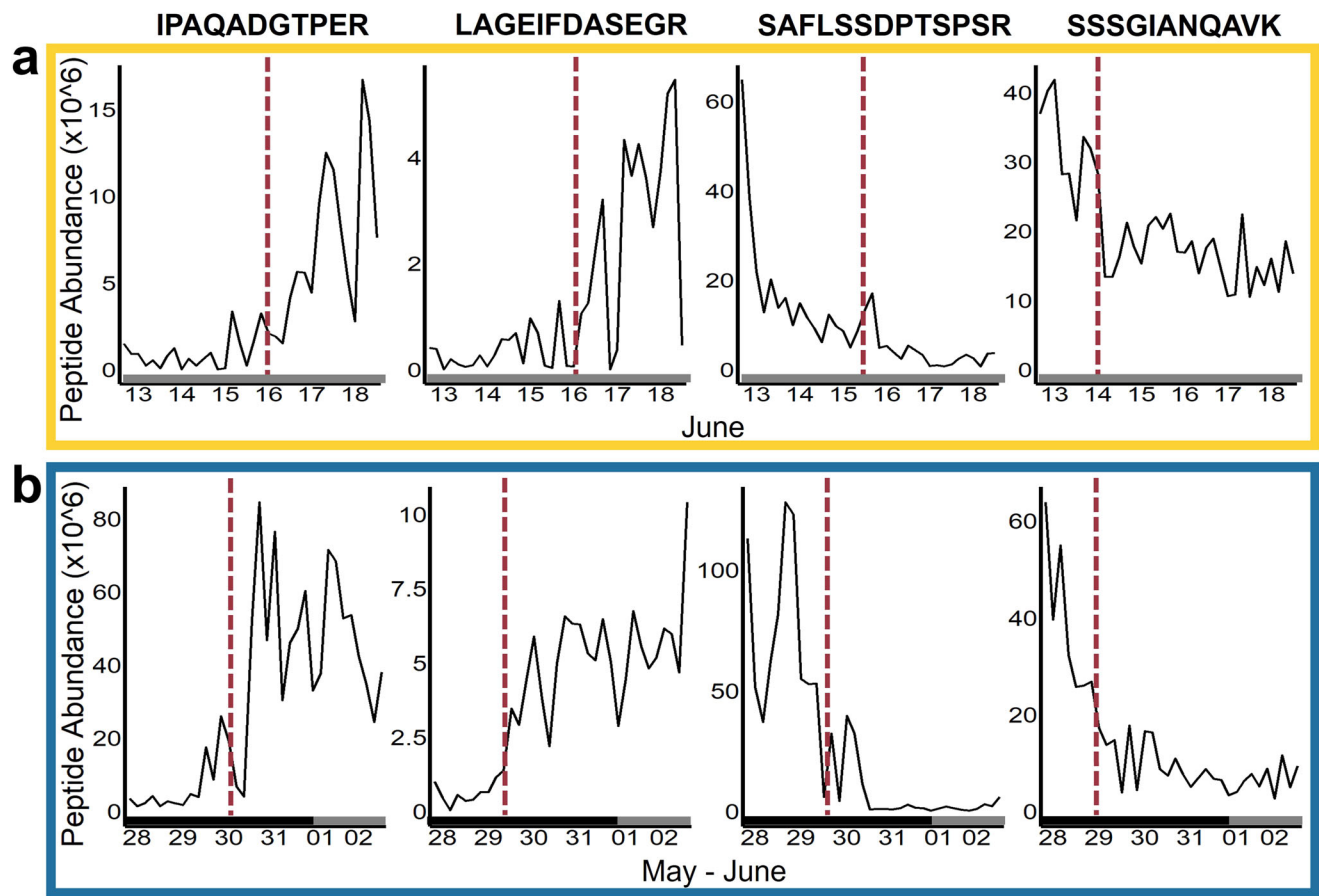


Fig. 5 | Peptides serve as reproducible predictive biomarkers of bloom onset across discovery and validation blooms. A representative subset of 4 of the 12 validated biomarkers' peptide abundances from (a) the discovery time series (yellow: June 12–18) and (b) the validation time series (blue: May 28 to June 02). State shift timing predicted by hidden Markov model indicated by dotted line. Source data are provided as a Source Data file.

Table 1 | Biomarker peptide annotations and details

Taxonomic Class	Peptide	Protein Function	Peptide Abundance Trend	Discovery Pre-bloom State Switch (hrs before BSD)	Validation Pre-bloom State Switch (hrs before BSD)
Gammaproteobacteria	APTPGQENVTK ^a	TonB-dependent receptor	+	66	54
	VVGLENQNPQNR ^a		+	62	54
	IPAQADGTPER	TonB-dependent receptor	+	42	34
	TAIPSYNVAR	TonB-dependent receptor	+	54	54
	LAGEIFDASEGR	30S ribosomal protein S7	+	46	58
	AAEANASPAVK	Bacteriorhodopsin-like protein	-	58	50
	SAFLSSDPTSPSR [^]	Family 16 glycosylhydrolase (GH16)	-	50	34
	VLNGNSGTVLGLNSR [^]		-	46	34
Flavobacteriia	SSSGIANQAVK	TonB-dependent receptor	-	102	66
	STPSVQINR	TonB-dependent receptor plug domain-containing protein	-	54	26
Verrucomicrobiae	AASTEAVNNAK	Prepilin-type N-terminal cleavage/ methylation domain-containing protein	+	78	30
	LQATGQLENTAR	50S ribosomal protein L29	+	66	34

^a,[^] Peptides map back to the same protein in metagenomic-generated protein sequence database. Successful candidate peptides are detailed with taxonomic class, protein of origin function, temporal trend of the quantified peptide abundance (increase as +, decrease as -), and timing of HMM state switch indicated as number of hours prior to BSD in the discovery and validation phytoplankton bloom events.

Since the DDA MS analysis of site- and time-matched samples provided peptide fragmentation spectra and their corresponding high mass accuracy precursor mass-to-charge ratios, traditional peptide-spectrum match search algorithms were applied to identify all detectable peptide candidates corresponding to the immense metagenomic protein sequence database. This information was then utilized to reduce the metagenomic protein sequence database by 99%. By doing the database reduction using DDA data that is distinct from the DIA data used in the discovery Phases, this approach avoids potential biases incurred by double-dipping methods that assess the value of signals of interest from the same data used to identify those signals, a widespread occurrence among other metaproteomic pipelines^{32,39,40}. This approach allowed for confident detection and quantification of DIA peptide signals in an unbiased, statistically robust manner, while accurately controlling the false discovery rate (FDR) (Supplementary Fig. 3). Presently, mass spectrometry analysis can pose a challenge for researchers unfamiliar with the field; however, as metaproteomic studies gain traction, such methods are becoming more accessible and affordable⁴⁰.

Due to the discovery-based nature of this analysis and the innate complexity of the marine microbiome, a combined approach of empirical and taxonomy-specific analyses for biomarker identification was advantageous, as neither individual analysis captured all peptides exhibiting biomarker potential and other ecosystems may exhibit different dynamics than those captured here. That said, many of the biomarker candidates were identified in both peptide selection methods in Phase 1 (88 peptides, 28%; Supplementary Table 1; Fig. 3). While Phase 1B (taxonomic correlation analysis) yielded over three times as many candidate peptides as Phase 1A (data dimensionality reduction), each method discovered a similar number of non-reverting peptides in Phase 2 (Phase 1A = 181; Phase 1B = 221). Limited sample quantities for downstream analyses restricted our validation screen to only a subset of peptides (87 of the 314; 28%) for targeted analysis; however, we posit that all 314 biomarkers identified in the discovery analysis have strong potential for development as predictive signals for HAB onset.

Although all samples were collected from the same coastal site, the two blooms began -17 days apart (Fig. 1a) with significantly different chemical landscapes represented by the varied nutrient profiles characterizing the water column prior to each bloom (Fig. 1h, i)³⁰. This pronounced difference between the conditions present during sample collection was seen as a benefit to the study as any bloom biomarker selected for real-world application would need to be consistent across diverse HAB systems and resilient to any external influences from potential perturbations. We acknowledge that despite finding predictive biomarkers, we have not revealed a mechanistic understanding or causal factors that lead to these detections. Further research is needed to identify mechanisms that lead to these detections.

Despite the contrasting temporal nutrient profiles, consistent microbial peptide signals were observed preceding both HABs (Fig. 5a, b; Supplementary Fig. 10). This observation underscores the microbiome's potential for revealing key changes in the complex chemical matrix of the ecosystem through highly specific biological responses independent of environmental conditions. Additionally, the timing of the HMM state shifts identified for all the peptides compared with the timing of the BSD for the two blooms revealed that a change in the collective functions of the bacterial microbiome occurred within 24–48 hours prior to bloom onset (Fig. 5a, b; Table 1). Here, we statistically identify and isolate the critical time frame in pre-bloom dynamics for future research and funding mechanisms to focus on.

The class-level bacterial structure of the microbiome in the pre-bloom discovery phase mirrors previously documented marine microbiomes surrounding phytoplankton bloom events^{27,41,42}. The classes dominating our final biomarker screen included commonly occurring

Gammaproteobacteria⁴³, Flavobacteriia^{44,45}, and Verrucomicrobiae^{46,47} (Table 1). The shared taxonomic profiles across blooms increase the likelihood that these peptides will be generalizable to other marine phytoplankton blooms and suggest the presence of keystone taxa that precede HAB formation. These classes all yielded detectable, quantifiable, and consistent biomarker peptides across both blooms (Fig. 5). A protein BLAST⁴⁸ analysis of each of the 12 validated peptides revealed that the active functionality represented by our biomarker panel was driven by general cellular growth and directed responses to chemical conditions of the pre-bloom ecosystem. Two of the biomarker peptides from Gammaproteobacteria uniquely map to the protein family 16 glycosylhydrolase (GH16), a class of carbohydrate-active enzymes (CAZymes) involved in the degradation of polysaccharides, the dominant molecules in transparent exopolymer particles (TEP) released by bloom forming algae^{49,50}. Another Gammaproteobacteria peptide that decreased as time approached the BSD was the light-driven proton pump bacteriorhodopsin-like protein⁵¹. A reduction in active light-derived energy processes, despite consistent light exposure (Supplementary Data 1), may indicate a switch to heterotrophic feeding strategies^{52,53}. These Gammaproteobacteria peptides may also reflect an active change in metabolism to employ general strategies for combating resource limitation and potential cell death, both conditions suggested by the overall downward trend of total peptides affiliated with the Gammaproteobacteria class (Fig. 4c). From Verrucomicrobiae, biomarker peptides mapped to the proteins 50S ribosomal protein L29 and a putative major type IV pilin protein (Prepilin-type N-terminal cleavage/methylation domain containing protein). We would expect an upregulation of protein translation-associated pathways, such as ribosomal proteins, to support the pre-bloom growth observed in Verrucomicrobiae⁵⁴ (Supplementary Data 9). Type IV pili are functionally diverse, playing a role in processes ranging from motility and biofilm formation to DNA uptake, protein secretion, and adherence^{55,56}, potentially allowing them to serve as a useful wide-ranging biomarker due to the prevalence of Verrucomicrobiae in marine ecosystems⁴⁷. All biomarkers exhibited a statistically significant, non-reverting state shift in abundance (Supplementary Fig. 7–8), allowing for the possibility of further development as a rapid assessment tool that can differentiate normal and pre-bloom microbiome states using a relative change in abundance threshold²³.

While a diverse screen is advantageous for capturing the heterogeneity of potential HAB systems, having multiple biomarkers that respond differently adds valuable redundancy, thus enhancing the overall reliability of the detection system⁵⁷. In our final peptide panel, 6 of the 12 biomarker peptides were identified as TonB-dependent receptors (TBDRs) (Table 1); TBDRs are components of a critical energy-dependent transport system for acquiring scarce vital resources⁵⁸. TonB systems are ubiquitous and highly expressed in gram-negative bacteria^{14,59,60}, and individual TBDRs exhibit high selectivity in substrate uptake, with targeted compounds including (but not limited to) iron⁶¹, vitamin B₁₂⁶², carbohydrates⁶³, lignans⁶⁴, and nickel chelates⁶⁵. While the high diversity of TBDRs makes it difficult to elucidate their specific functional role in this ecosystem, the contrasting patterns between the TBDRs in bacterial classes detected here (i.e., increasing abundance in Gammaproteobacteria vs. decreasing abundance in Flavobacteriia) suggests that distinct chemical signals in the water column may be preferentially beneficial to specific organisms within the microbiome^{45,66,67}. Previous studies have demonstrated that both Gammaproteobacteria and Flavobacteriia utilize TBDRs in response to algal blooms, specifically to consume complex organic molecules (i.e., polysaccharides) released throughout bloom progression by algae as they increase their metabolic productivity^{43,44,68,69}. Here we show, using in situ evidence that TBDR proteins may be active prior to bloom formation. Because TonB systems are a metabolically expensive transport method, the cosmopolitan frontline response seen here suggests TBDRs equate to an essential and highly sensitive

indicator of changes in water chemistry precluding HAB initiation, posing an ideal target for the development of a forecasting tool.

Techniques for bloom monitoring and forecasting that utilize water chemistry⁷⁰, fluid dynamics⁷¹, satellites⁷², machine learning⁷³, and even the microbiome¹⁹ have been proposed; however, to date, no reliable predictor of HABs exists. Current monitoring techniques (i.e., in situ imaging, toxin analysis) require a bloom to reach a critical mass or produce dangerous toxins before HAB detection is confirmed; in contrast, the peptide biomarkers detailed here undergo a detectable and significant shift *prior* to bloom development thereby allowing sufficient time to alert critical reporting agencies of the oncoming dangers of bloom development prior to public exposure and safety issues. Additionally, the unique state shift occurring in the biomarker peptide abundances lends itself to the development of an assay in which the biomarkers act as an on/off switch for determining bloom initiation, eliminating the need for high-resolution monitoring and cutting development costs for stakeholders invested in using the tool.

Further study is needed in both oceanic and coastal marine ecosystems to validate if these findings are translatable to blooms of other planktonic types, including cyanobacteria and toxin-producing blooms. While both blooms reported here were documented to affect the greater East Sound area, characterization of the magnitude of the bloom required to elicit a detectable response in the preceding microbiome shift would be beneficial for determining the robustness of these biomarkers across bloom systems of varying duration and intensity. Because marine microbiomes globally share similarities in taxonomic structure^{15–17}, and bacteria commonly associated with HAB-forming phytoplankton exhibit a functional redundancy that creates overlap in the metabolic potentials of unique taxa^{28,74}, we hypothesize that the bacterial microbiome could generate biomarkers applicable for blooms of various types (diatom, dinoflagellate, raphidophyte, toxin vs non-toxin producing, etc.). This research highlights both the potential of the microbiome for generating intracellular predictive signals as well as the potential for dissolved metabolomics analyses to elucidate extracellular biochemical signals in the water that may be driving bloom formation. While capturing and characterizing a bloom system prior to onset remains difficult⁷⁵, isolating this critical time frame in future studies could advance our understanding of what triggers a bloom and how community structure is impacted by environmental factors prompting bloom initiation.

Methods

Sample collection and water column analysis

Sample collection methods are detailed in Nunn et. al., 2024³⁰. Briefly, field sampling was conducted in East Sound, WA, USA from May 27 to June 18, 2021, by pumping water from 100 m offshore at a constant depth of 2 m and flow rate of 5 L min⁻¹ to obtain 20 L. From the 20 L, multiple peristaltic pumps were set up to simultaneously filter between 500 – 2000 mL total volume per sample, averaging 1 L per microbiome filter. Whole water was filtered through a 200 µm nylon mesh followed by three subsequent filters at 100 µm, 10 µm, 1 µm before final collection onto a 0.22 µm polyethersulfone (PES) 47 mm filter. Final metaproteomic samples were stored long term in -80 °C on 0.22 µm polyethersulfone (PES) filters. Additional samples collected included nutrients, flow cytometry, prokaryotic DNA/proteomics, eukaryotic DNA/RNA/proteomics, and dissolved metabolites. Water chemistry was also monitored at the collection site with a YSI Exo1 Sonde probe capturing chlorophyll *a* concentration, salinity, temperature, dissolved oxygen, and pH every 10 minutes.

Metagenome assembly and database generation

DNA extraction was performed according to methods detailed in Nunn et al.³⁰ on samples from all time points collected. DNA from all 132 samples were combined to generate a pooled sample with equal amounts of DNA (50 ng) from each sample for whole genome

sequencing (WGS) using 150 bp paired-end Illumina sequencing (NextSeq 500 Mid Output v2 Kit - 300 cycle) through the Northwest Genomics Center (NWGC) at the University of Washington. DNA library preparation, quality controls, and WGS were performed by the NWGC. The resulting FASTQs were assembled into a metagenome using MEGAHIT⁷⁶. Samtools⁷⁷ 1.14 was used with Python⁷⁸ 3.7.7 to filter only mapped reads, and Prodigal⁷⁹ 2.6.3 was used to convert contigs into a FASTA file consisting of 3,457,818 proteins (PRIDE identifier PXD052873).

Proteomic sample processing

Samples collected for microbiome proteomic analysis (collected on 0.22 µm PES) were manually agitated within a 3×4 cm Ziplock baggie with 100 µl S-Trap solubilization buffer (5% SDS, 50 mM TEAB, 2 mM MgCl₂, 1X HALT protease and phosphatase inhibitors in nanopure water) to ensure equal spreading of the buffer across the filter surface. The cell lysate was collected via pipette into a 2 mL microfuge tube. This was followed by an additional rinse of the filter with 100 µl nanopure water, followed by manual agitation and immediate transfer of the solution to the pooled fractions in the 2 mL tube. Two additional rinses of the filter with 100 µl nanopure water without manual agitation were performed and the solution added to the pool. The resulting 400 µl whole cell solution was then sonicated (Branson 250 Sonifier; 20 kHz, 3 × 10 s on ice). Samples were then evaporated in a SpeedVac (ThermoFisher Scientific SRF110) at room temperature to reduce the volume to 100 µL (5% SDS). Protein concentrations for each sample were quantified in technical triplicates using a Bicinchoninic Acid protein assay kit (Pierce Thermo Scientific) using a microplate reader. A total of 20 µg protein per sample was isolated for enzymatic digestion and MS-based identification. Enolase protein standard (0.16 µL of 100 ng µl⁻¹ enolase per 1 µg protein) was added to each sample to track digestion efficiency. Each sample (20 µg protein) was treated with benzonase to degrade DNA (0.5 µL 250 unit µl⁻¹ 10 minutes at 95 °C), and proteins were reduced with 20 mM dithiothreitol for 10 minutes at 60 °C followed by a four minute cool down to room temperature, alkylated with 40 mM iodoacetamide for 30 minutes in the dark, acidified to pH <2 (using 1.2% aqueous phosphoric acid), and then processed on an S-trap micro column according to manufacturer's recommendations (Protifi S-Trap micro protocol). Proteins were digested with Promega modified trypsin (2 µg for 1:10 ratio, 1 hour 37 °C; ThermoFisher Scientific). Resulting peptides were evaporated to dryness and resuspended in 2% acetonitrile (ACN), 0.1% formic acid with final concentration of 0.5 µg protein µl⁻¹. Prior to analysis on the mass spectrometer, each sample received 50 fmol µg⁻¹ protein of Pierce Retention Time Calibration (PRTC) as an external standard to monitor MS performance across experiments.

Proteomic mass spectrometry overview

All mass spectrometry analyses were performed on an Orbitrap Lumos Fusion Tribrid instrument with an inline EASY-nLC 1200 (ThermoFisher Scientific). Reverse phase chromatography was achieved using a manufactured PicoTip fused silica capillary column (75 µm i.d., 40 cm long) packed in-house with C18 beads (Dr. Maisch ReproSil-Pur; C18-Aq, 120 Å, 3 µm) and maintained at 50 °C. The analytical column was fitted with a 4 cm long, 150 µm i.d. in-house packed kasil-frit pre-column (Dr. Maisch ReproSil-Pur; C18-Aq, 120 Å, 3 µm). Peptides were eluted using an acidified (formic acid, 0.1% v/v) water-acetonitrile gradient (2–45% acetonitrile). Each sample was analyzed with data-dependent acquisition (DDA) and data-independent acquisition (DIA) in random order. Skyline was used to evaluate the consistency of the resulting peak areas and retention times of QC standards through all analyses⁸⁰. The DDA MS data are deposited to the Proteome Xchange Consortium via the PRIDE partner repository with the dataset identifier PXD052873. The DIA and PRM MS data are deposited to Panorama Public^{81,82}: (<https://panoramaweb.org/Orcas-Island-Time-Series.url>).

DDA analyses

Samples from June 12, 2021 (17:00) to June 18, 2021 (13:00) were analyzed on the mass spectrometer operating in DDA mode. The top 20 most intense ions were selected for MS2 acquisition from precursor ion scans of 400–1200 *m/z*. Centroid full MS resolution data was collected at 70,000 with AGC target of 1×10^6 and centroid MS2 data was collected at resolution of 35,000 with an AGC target of 5×10^4 . Dynamic exclusion was set to 10 seconds and +2, +3, +4 ions were selected for MS2. Comet 2022.01^{83,84} was used to search the DDA files; parameters included: reverse concatenated sequence database search, 10 ppm precursor mass tolerance, semi-tryptic specificity with 3 allowed missed cleavages, cysteine modification of 57 Da (resulting from the iodoacetamide) and modifications on methionine of 15.999 Da (oxidation). Percolator (version 3.02)^{85,86} was used to select peptides with 1% FDR. An open-source Nextflow workflow (<https://github.com/mriddle/nf-filter-fasta>) was used to reduce the metagenomic FASTA from 3,457,818 proteins into a smaller DDA-refined database representing all possible proteins from the peptides detected in the samples for proteomic interpretation. These peptides were mapped back to the full metagenomic FASTA to compile a database of all proteins with at least 1 peptide identified by the workflow. The resulting DDA-refined database contained 36,429 protein sequences (Supplementary Data 2).

DIA analyses

Samples from June 12, 2021 (17:00) to June 18, 2021 (13:00) were analyzed on the mass spectrometer using DIA mode. For each DIA MS experiment, 4 μ g digested peptides with 7 fmol PRTC were analyzed. Centroid full MS resolution data was collected at 120,000 and MS2 data was collected at a resolution of 15,000 with an AGC target of 4×10^5 across isolation windows of 8 *m/z*. Prosit⁸⁷ was used to process the DDA-refined database using an HCD intensity and ion retention time prediction model according to methods from Searle et. al., 2020⁸⁸. Six gas phase fractionations (GPFs) spanning 100 *m/z* for each analysis were completed with a narrow isolation window (4 *m/z*) to generate a chromatogram library using Encyclopedia⁸⁹ v. 1.4.01 following Pino et al., 2020⁹⁰. The 36 individual time series MS experiments were searched against the resulting chromatogram library using Encyclopedia v. 1.12.31. Encyclopedia-exported results were visualized using Skyline-daily version 22.2.0.527, including quality assurance that every peptide contained at least three quantifiable transitions and external standards (PRTC peptides) had a coefficient of variation below 20%. Quantitative values for each peptide were generated using Skyline, which calculated the area under the curve (AUC) for each peptide after total ion current (TIC) normalization using Encyclopedia-inferred peak boundaries.

Taxonomic and Functional Annotations

MetaGOMics³⁴ software (version 0.2.0; <https://www.yeastrc.org/metagomics>) identifies the lowest common ancestor on each peptide's taxonomic and functional Gene Ontology (GO) assignment using a provided site-specific metagenomic-generated protein sequence database. Here, we used the DDA-refined database as our background proteome and provided a complete list of all peptides identified at each timepoint with their respective peptide quantities (DIA-MS). The MetaGOMics reports were used to annotate peptides to their respective GO terms and taxonomic classes.

Framework for selecting peptides

Phase 1A: Data dimensionality reduction- PCA. PC1 was driven by missingness of data across the 36 timepoints and was not included in further exploratory analysis. PC2 and PC3 were separated by day of collection, identifying a critical time point of change. The top 10% of peptides driving variation across PC2 were selected as candidate peptides (Supplementary Data 4).

Phase 1B: Taxonomic trends in summed peptide abundances. Peptide abundances resulting from each class as annotated by MetaGOMics, were summed at each time point and plotted through time (Supplementary Data 5). A Spearman's correlation coefficient ρ (ρ) was then calculated (Supplementary Table 2; Supplementary Fig. 5) comparing the summed quantitative peptide abundance values for each class and ranked time. All peptides from the bacterial classes with a correlation coefficient $\rho \geq |0.5|$ were selected for further development and testing as biomarkers (i.e., Flavobacteriia, Gammaproteobacteria, Verrucomicrobiae, and Opitutae). Alphaproteobacteria was additionally added to the analysis as it represented the most abundant taxonomic class.

Phase 2: HMM State-shift time series analysis. A two-state hidden Markov model (HMM) was then fit to the quantitative values associated with each candidate peptide to identify those that undergo a non-reverting state shift >24 hours prior to the BSD (R package depmixS4⁹¹; Supplementary Fig. 7–8). The resulting probability matrix was used to determine the latent state of the observations, and the determined state outcomes were plotted through time. Of the 314 peptides determined to undergo a non-reverting state shift, 87 were selected for the next phase of biomarker development.

Phase 3: Matrix-Matched Calibration Curve (MMCC). To determine if the candidate peptides are quantitative by mass spectrometry, a matrix-matched calibration curve (MMCC) was created³⁵. MMCC analysis provides the user with the ability to determine a peptide's detectability within a complex matrix and whether the peptide yields a linear response when analyzed with targeted parallel reaction monitoring (PRM) methods in a dilution series. A pooled sample consisting of 2 μ g aliquots of digested peptides from 36 samples from our time series served as a representative water microbiome. This microbiome sample was serially diluted with a matched matrix of peptides resulting from the marine bacterium *Colwellia psychrerythraea* 34H⁹². The resulting dilution series consisted of six microbiome-to-matrix ratios that spanned the expected dynamic range of targeted peptides: 0%, 0.3%, 3%, 50%, 70%, 100% microbiome.

PRM analyses on MMCC

The targeted peptide list was generated using the Skyline analysis of the DIA data. Peptides were trapped and separated on a near-identical chromatography setup detailed earlier with the exception of the analytical column length (30 cm). Mass spectrometry experiments were completed with 2 μ g digested peptides with 150 fmol PRTC. Peptides were eluted using the same gradient as DDA/DIA experiments followed by a 30 minute wash with 80% ACN and 10 minute re-equilibration with 3% ACN. Targeted ions were selected for MS2 acquisition from precursor ion scans of 400–1000 *m/z* with an isolation window of 1.6 *m/z*, flow rate of 300 nL/min, and resolution of 30,000. Centroid full MS and MS2 data was collected with an AGC target of 4×10^5 . Data was analyzed with Skyline⁸⁰ software to identify peptide peaks and chromatographic time boundaries. Skyline-calculated peptide quantification from AUC were exported and analyzed with R following Pino et al.³⁵. A total of 27 peptides were identified to yield linear responses via the MMCC with resulting limit of detection (LOD) > 1e6.

Biomarker Peptide Validation

To validate the final 27 candidate peptides as predictive HAB biomarkers, PRM MS analyses were performed on the set of samples collected from May 28, 2021 (01:00) to June 3, 2021 (01:00) using the same target peptide list. MS experiments were completed on 2 μ g peptides with 150 fmol PRTC external standard. Skyline (version 23.0.9.187) was used to visualize the peptide peak boundaries, which were visualized and manually adjusted based on the parameters established by the MMCC methods. The resulting AUC report was

exported from Skyline and evaluated in R to analyze the temporal trends for each peptide. A two-state HMM was fit to each peptide to select for those that undergo a non-reverting state shift, resulting in the final panel of 12 successful biomarker peptides. A Spearman correlation analysis was performed on the 12 peptides to compare peptide abundances across the discovery and validation data sets and assess for consistency of pre-bloom temporal trends. Taxonomic and functional annotations were assigned to each of the 12 peptides by first mapping each peptide back to its respective parent proteins within the DDA-refined database. Subsequently, a standard protein search was performed against the publicly available NR database using the Basic Local Alignment Search Tool (BLAST; 2/6/2024) (Table 1).

Statistics and Reproducibility

Sample analyses on the MS were randomized to reduce batch effects and quality control (QC) peptide mixtures were analyzed every sixth injection to monitor chromatography and MS sensitivity. Quantitative peptide AUC values were first median normalized and log transformed prior to performing the PCA. A Mann-Whitney test was used to test for significance of the PCA results between the first and second half of the timeseries. For testing correlation strength throughout the study, a Spearman's rank-order correlation test was used to calculate Spearman's rho (ρ) and significance was assessed by calculating a two-tailed p-value. The two-state hidden Markov model (HMM) used throughout the study was fit using the depmixS4 package. Time series data was log transformed prior to model fitting. The model was specified to have a Gaussian distribution and transition probabilities were assumed to be stationary. Model parameters were estimated using an Expectation-Maximization (EM) algorithm, with a random seed set to ensure reproducibility. Data are presented as either raw values or the mean $\pm$ SD of at least duplicate measures. No data were excluded from the analyses. All statistical analyses were performed using R version 4.3.2.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

The bacterial metagenome reference sequences generated in this study are available under the NCBI BioSample accession [SAMN41720822](https://www.ncbi.nlm.nih.gov/biosample/SAMN41720822) associated with the BioProject [PRJNA1130528](https://www.ncbi.nlm.nih.gov/bioproject/PRJNA1130528). The assembled FASTA generated using MEGAHIT is available with the PRIDE identifier [PXD052873](https://www.ebi.ac.uk/PRIDE/peptide-spectra/PXD052873). The DDA MS data are deposited to the Proteome Xchange Consortium via the PRIDE partner repository with the dataset identifier [PXD052873](https://www.ebi.ac.uk/PRIDE/peptide-spectra/PXD052873). The DIA and PRM MS data are deposited to Panorama Public^{81,82}: (<https://panoramaweb.org/Orcas-Island-Time-Series.url>). The datasets generated in this study are provided on figshare (<https://doi.org/10.6084/m9.figshare.27166347>). Further details on the datasets are provided in the Supplementary Information. Source data are provided with this paper.

Code availability

All code was generated using R version 4.3.2⁹³. Code and scripts used for this study are available at: <https://github.com/Nunn-Lab/Publication-2021-Orcas-Island-Time-Series>⁹⁴.

References

1. What is a harmful algal bloom? | National Oceanic and Atmospheric Administration. <https://www.noaa.gov/what-is-harmful-algal-bloom>.
2. Berdalet, E. et al. Marine harmful algal blooms, human health and wellbeing: challenges and opportunities in the 21st century. *J. Mar. Biol. Assoc. U. K.* **96**, 61–91 (2016).
3. Anderson, D. M. et al. Marine harmful algal blooms (HABs) in the United States: History, current status and future trends. *Harmful Algae* **102**, 101975 (2021).
4. Brooks, B. W. et al. Are harmful algal blooms becoming the greatest inland water quality threat to public health and aquatic ecosystems? *Environ. Toxicol. Chem.* **35**, 6–13 (2016).
5. Turley, B. D., Karnauskas, M., Campbell, M. D., Hanisko, D. S. & Kelble, C. R. Relationships between blooms of *Karenia brevis* and hypoxia across the West Florida Shelf. *Harmful Algae* **114**, 102223 (2022).
6. Rolton, A. et al. Effects of Harmful Algal Blooms on Fish and Shellfish Species: A Case Study of New Zealand in a Changing Environment. *Toxins* **14**, 341 (2022).
7. Glibert, P. M. & Burkholder, J. M. Causes of Harmful Algal Blooms. in *Harmful Algal Blooms* 1–38 (John Wiley & Sons, Ltd, 2018).
8. Barton, A. D., Irwin, A. J., Finkel, Z. V. & Stock, C. A. Anthropogenic climate change drives shift and shuffle in North Atlantic phytoplankton communities. *Proc. Natl Acad. Sci. USA.* **113**, 2964–2969 (2016).
9. Mayali, X. Editorial: Metabolic Interactions Between Bacteria and Phytoplankton. *Front. Microbiol.* **9**, 727 (2018).
10. Seymour, J. R., Amin, S. A., Raina, J.-B. & Stocker, R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nat. Microbiol.* **2**, 1–12 (2017).
11. Muratore, D. et al. Complex marine microbial communities partition metabolism of scarce resources over the diel cycle. *Nat. Ecol. Evol.* **6**, 218–229 (2022).
12. Kuhlisch, C., Shemi, A., Barak-Gavish, N., Schatz, D. & Vardi, A. Algal blooms in the ocean: hot spots for chemically mediated microbial interactions. *Nat. Rev. Microbiol.* **22**, 138–154 (2024).
13. Coyne, K. J., Wang, Y. & Johnson, G. Algicidal Bacteria: A Review of Current Knowledge and Applications to Control Harmful Algal Blooms. *Front. Microbiol.* **13**, 871177 (2022).
14. Li, D. X. et al. Metaproteomics reveals major microbial players and their metabolic activities during the blooming period of a marine dinoflagellate *Prorocentrum donghaiense*. *Environ. Microbiol.* **20**, 632–644 (2018).
15. Deng, Y., Wang, K., Hu, Z., Hu, Q. & Tang, Y. Z. Identification and implications of a core bacterial microbiome in 19 clonal cultures laboratory-reared for months to years of the cosmopolitan dinoflagellate *Karlodinium veneticum*. *Front. Microbiol.* **13**, 967610 (2022).
16. Fortin, S. G., Song, B., Anderson, I. C. & Reece, K. S. Blooms of the harmful algae *Margalefidinium polykrikoides* and *Alexandrium monilatum* alter the York River Estuary microbiome. *Harmful Algae* **114**, 102216 (2022).
17. Hattenrath-Lehmann, T. K. & Gobler, C. J. Identification of unique microbiomes associated with harmful algal blooms caused by *Alexandrium fundyense* and *Dinophysis acuminata*. *Harmful Algae* **68**, 17–30 (2017).
18. Grossart, H.-P., Levold, F., Allgaier, M., Simon, M. & Brinkhoff, T. Marine diatom species harbour distinct bacterial communities. *Environ. Microbiol.* **7**, 860–873 (2005).
19. Gajardo, G. et al. The holobiome of marine harmful algal blooms (HABs): A novel ecosystem-based approach for implementing predictive capabilities and managing decisions. *Environ. Sci. Policy* **143**, 44–54 (2023).
20. Calado, C. R. C. Antigenic and conserved peptides from diverse *Helicobacter pylori* antigens. *Biotechnol. Lett.* **44**, 535–545 (2022).
21. Maier, T. et al. Quantification of mRNA and protein and integration with protein turnover in a bacterium. *Mol. Syst. Biol.* **7**, 511 (2011).
22. Brunner, A.-D. et al. Ultra-high sensitivity mass spectrometry quantifies single-cell proteome changes upon perturbation. *Mol. Syst. Biol.* **18**, e10798 (2022).

23. Pandey, S., Malviya, G. & Chottova Dvorakova, M. Role of Peptides in Diagnostics. *Int. J. Mol. Sci.* **22**, 8828 (2021).
24. Agarwal, D. K. et al. Highly sensitive and ultra-rapid antigen-based detection of SARS-CoV-2 using nanomechanical sensor platform. *Biosens. Bioelectron.* **195**, 113647 (2022).
25. Wang, Y. et al. Advances and development in sampling techniques for marine water resources: a comprehensive review. *Front. Mar. Sci.* **11**, 1365019 (2024).
26. Keeling, P. J. et al. The Marine Microbial Eukaryote Transcriptome Sequencing Project (MMETSP): Illuminating the Functional Diversity of Eukaryotic Life in the Oceans through Transcriptome Sequencing. *PLOS Biol.* **12**, e1001889 (2014).
27. Ferrer-González, F. X. et al. Resource partitioning of phytoplankton metabolites that support bacterial heterotrophy. *ISME J.* **15**, 762–773 (2021).
28. Li, D. et al. Quantifying functional redundancy in polysaccharide-degrading prokaryotic communities. *Microbiome* **12**, 120 (2024).
29. Li, X.-Y. et al. Marked shifts of harmful algal blooms in the Bohai Sea linked with combined impacts of environmental changes. *Harmful Algae* **121**, 102370 (2023).
30. Nunn, B. L. et al. Microbial Metagenomes Across a Complete Phytoplankton Bloom Cycle: High-Resolution Sampling Every 4 Hours Over 22 Days. *Sci. Data* **11**, 1270 (2024).
31. Brody, S. R., Lozier, M. S. & Dunne, J. P. A comparison of methods to determine phytoplankton bloom initiation. *J. Geophys. Res. Oceans* **118**, 2345–2357 (2013).
32. Everett, L. J., Bierl, C. & Master, S. R. Unbiased Statistical Analysis for Multi-Stage Proteomic Search Strategies. *J. Proteome Res.* **9**, 700–707 (2010).
33. Bern, M. & Kil, Y. J. Comment on “Unbiased statistical analysis for multi-stage proteomic search strategies”. *J. Proteome Res.* **10**, 2123–2127 (2011).
34. Riffle, M. et al. MetaGOmics: A Web-Based Tool for Peptide-Centric Functional and Taxonomic Analysis of Metaproteomics Data. *Proteomes* **6**, 2 (2017).
35. Pino, L. K. et al. Matrix-Matched Calibration Curves for Assessing Analytical Figures of Merit in Quantitative Proteomics. *J. Proteome Res.* **19**, 1147–1153 (2020).
36. Timmins-Schiffman, E., Mikan, M. P., Ting, Y. S., Harvey, H. R. & Nunn, B. L. MS analysis of a dilution series of bacteria: phytoplankton to improve detection of low abundance bacterial peptides. *Sci. Rep.* **8**, 9276 (2018).
37. Ahern, O. M., Whittaker, K. A., Williams, T. C., Hunt, D. E. & Ryneason, T. A. Host genotype structures the microbiome of a globally dispersed marine phytoplankton. *Proc. Natl Acad. Sci. USA* **118**, e2105207118 (2021).
38. Gregor, J. & Maršálek, B. Freshwater phytoplankton quantification by chlorophyll a: a comparative study of in vitro, in vivo and in situ methods. *Water Res* **38**, 517–522 (2004).
39. Ball, T. M., Squeglia, L. M., Tapert, S. F. & Paulus, M. P. Double Dipping in Machine Learning: Problems and Solutions. *Biol. Psychiatry Cogn. Neurosci. Neuroimaging* **5**, 261–263 (2020).
40. Salvato, F., Hettich, R. L. & Kleiner, M. Five key aspects of meta-proteomics as a tool to understand functional interactions in host-associated microbiomes. *PLoS Pathog.* **17**, e1009245 (2021).
41. Patin, N. V. et al. Microbial and chemical dynamics of a toxic dinoflagellate bloom. *PeerJ* **8**, e9493 (2020).
42. Lindh, M. V. et al. Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. *Environ. Microbiol.* **17**, 2459–2476 (2015).
43. Francis, B., Urich, T., Mikolasch, A., Teeling, H. & Amann, R. North Sea spring bloom-associated Gammaproteobacteria fill diverse heterotrophic niches. *Environ. Microbiome* **16**, 15 (2021).
44. Williams, T. J. et al. The role of planktonic Flavobacteria in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environ. Microbiol.* **15**, 1302–1317 (2013).
45. Teeling, H. et al. Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife* **5**, e11888 (2016).
46. Sichert, A. et al. Verrucomicrobia use hundreds of enzymes to digest the algal polysaccharide fucoidan. *Nat. Microbiol.* **5**, 1026–1039 (2020).
47. Orellana, L. H. et al. Verrucomicrobiota are specialist consumers of sulfated methyl pentoses during diatom blooms. *ISME J.* **16**, 630–641 (2022).
48. Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **215**, 403–410 (1990).
49. Viborg, A. H. et al. A subfamily roadmap of the evolutionarily diverse glycoside hydrolase family 16 (GH16). *J. Biol. Chem.* **294**, 15973–15986 (2019).
50. Vidal-Melgosa, S. et al. Diatom fucan polysaccharide precipitates carbon during algal blooms. *Nat. Commun.* **12**, 1150 (2021).
51. Bèjà, O. et al. Bacterial Rhodopsin: Evidence for a New Type of Phototrophy in the Sea. *Science* **289**, 1902–1906 (2000).
52. Bar-Shalom, R. et al. Rhodopsin-mediated nutrient uptake by cultivated photoheterotrophic Verrucomicrobiota. *ISME J.* **17**, 1063–1073 (2023).
53. Guo, Z., Zhang, H. & Lin, S. Light-Promoted Rhodopsin Expression and Starvation Survival in the Marine Dinoflagellate *Oxyrrhis marina*. *PLoS ONE* **9**, e11494 (2014).
54. Scott, M., Gunderson, C. W., Mateescu, E. M., Zhang, Z. & Hwa, T. Interdependence of cell growth and gene expression: origins and consequences. *Science* **330**, 1099–1102 (2010).
55. Ellison, C. K., Whitfield, G. B. & Brun, Y. V. Type IV Pili: dynamic bacterial nanomachines. *FEMS Microbiol. Rev.* **46**, fuab053 (2022).
56. Bünger, W., Jiang, X., Müller, J., Hurek, T. & Reinhold-Hurek, B. Novel cultivated endophytic Verrucomicrobia reveal deep-rooting traits of bacteria to associate with plants. *Sci. Rep.* **10**, 8692 (2020).
57. Ou, F.-S., Michiels, S., Shyr, Y., Adjei, A. A. & Oberg, A. L. Biomarker Discovery and Validation: Statistical Considerations. *J. Thorac. Oncol. Publ. Int. Assoc. Study Lung Cancer* **16**, 537–545 (2021).
58. Silale, A. & Berg, B. van den. TonB-Dependent Transport Across the Bacterial Outer Membrane. *Annu. Rev. Microbiol.* **77**, 67–88 (2023).
59. Tang, K., Jiao, N., Liu, K., Zhang, Y. & Li, S. Distribution and Functions of TonB-Dependent Transporters in Marine Bacteria and Environments: Implications for Dissolved Organic Matter Utilization. *PLOS ONE* **7**, e41204 (2012).
60. Morris, R. M. et al. Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J.* **4**, 673–685 (2010).
61. Pawelek, P. D. et al. Structure of TonB in Complex with FhuA, E. coli Outer Membrane Receptor. *Science* **312**, 1399–1402 (2006).
62. Shultz, D. D., Purdy, M. D., Banchs, C. N. & Wiener, M. C. Outer Membrane Active Transport: Structure of the BtuB: TonB Complex. *Science* **312**, 1396–1399 (2006).
63. Blavillain, S. et al. Plant Carbohydrate Scavenging through TonB-Dependent Receptors: A Feature Shared by Phytopathogenic and Aquatic Bacteria. *PLoS ONE* **2**, e224 (2007).
64. Fujita, M. et al. A TonB-dependent receptor constitutes the outer membrane transport system for a lignin-derived aromatic compound. *Commun. Biol.* **2**, 1–10 (2019).
65. Schauer, K., Gouget, B., Carrière, M., Labigne, A. & De Reuse, H. Novel nickel transport mechanism across the bacterial outer membrane energized by the TonB/ExbB/ExbD machinery. *Mol. Microbiol.* **63**, 1054–1068 (2007).

66. Russo, D. A., Couto, N., Beckerman, A. P. & Pandhal, J. A Metaproteomic Analysis of the Response of a Freshwater Microbial Community under Nutrient Enrichment. *Front. Microbiol.* **7**, 1172 (2016).
67. Pollet, R. M., Martin, L. M. & Koropatkin, N. M. TonB-dependent transporters in the Bacteroidetes: Unique domain structures and potential functions. *Mol. Microbiol.* **115**, 490–501 (2021).
68. Francis, T. B. et al. Changing expression patterns of TonB-dependent transporters suggest shifts in polysaccharide consumption over the course of a spring phytoplankton bloom. *ISME J.* **15**, 2336–2350 (2021).
69. Dong, H.-P., Hong, Y.-G., Lu, S. & Xie, L.-Y. Metaproteomics reveals the major microbial players and their biogeochemical functions in a productive coastal system in the northern South China Sea. *Environ. Microbiol. Rep.* **6**, 683–695 (2014).
70. Marrone, B. L., Banerjee, S., Talapatra, A., Gonzalez-Esquer, C. R. & Pilania, G. Toward a Predictive Understanding of Cyanobacterial Harmful Algal Blooms through AI Integration of Physical, Chemical, and Biological Data. *ACS EST Water* **4**, 844–858 (2024).
71. Zheng, L. et al. Prediction of harmful algal blooms in large water bodies using the combined EFDC and LSTM models. *J. Environ. Manag.* **295**, 113060 (2021).
72. Lin, J., Miller, P. I., Jönsson, B. F. & Bedington, M. Early warning of harmful algal bloom risk using satellite ocean color and Lagrangian particle trajectories. *Front. Mar. Sci.* **8**, 736262 (2021).
73. Lee, D. et al. Integrated explainable deep learning prediction of harmful algal blooms. *Technol. Forecast. Soc. Change* **185**, 122046 (2022).
74. Tian, L. et al. Deciphering functional redundancy in the human microbiome. *Nat. Commun.* **11**, 6217 (2020).
75. Ralston, D. K. & Moore, S. K. Modeling harmful algal blooms in a changing climate. *Harmful Algae* **91**, 101729 (2020).
76. Li, D., Liu, C.-M., Luo, R., Sadakane, K. & Lam, T.-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* **31**, 1674–1676 (2015).
77. Danecek, P. et al. Twelve years of SAMtools and BCFtools. *Giga-Science* **10**, giab008 (2021).
78. Van Rossum, G. & Drake, F. L. *Python 3 Reference Manual*. (CreateSpace, Scotts Valley, CA, 2009).
79. Hyatt, D. et al. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinforma.* **11**, 119 (2010).
80. MacLean, B. et al. Skyline: an open source document editor for creating and analyzing targeted proteomics experiments. *Bioinforma. Oxf. Engl.* **26**, 966–968 (2010).
81. Sharma, V. et al. Panorama: A Targeted Proteomics Knowledge Base. *J. Proteome Res.* **13**, 4205–4210 (2014).
82. Sharma, V. et al. Panorama Public: A Public Repository for Quantitative Data Sets Processed in Skyline. *Mol. Cell. Proteom. MCP* **17**, 1239–1244 (2018).
83. Eng, J. K., Jahan, T. A. & Hoopmann, M. R. Comet: an open-source MS/MS sequence database search tool. *Proteomics* **13**, 22–24 (2013).
84. Eng, J. K. et al. A deeper look into Comet – implementation and features. *J. Am. Soc. Mass Spectrom.* **26**, 1865–1874 (2015).
85. Käll, L., Canterbury, J. D., Weston, J., Noble, W. S. & MacCoss, M. J. Semi-supervised learning for peptide identification from shotgun proteomics datasets. *Nat. Methods* **4**, 923–925 (2007).
86. The, M., MacCoss, M. J., Noble, W. S. & Käll, L. Fast and Accurate Protein False Discovery Rates on Large-Scale Proteomics Data Sets with Percolator 3.0. *J. Am. Soc. Mass Spectrom.* **27**, 1719–1727 (2016).
87. Gessulat, S. et al. Prosit: proteome-wide prediction of peptide tandem mass spectra by deep learning. *Nat. Methods* **16**, 509–518 (2019).
88. Searle, B. C. et al. Generating high quality libraries for DIA MS with empirically corrected peptide predictions. *Nat. Commun.* **11**, 1548 (2020).
89. Searle, B. C. et al. Chromatogram libraries improve peptide detection and quantification by data independent acquisition mass spectrometry. *Nat. Commun.* **9**, 5128 (2018).
90. Pino, L. K., Just, S. C., MacCoss, M. J. & Searle, B. C. Acquiring and Analyzing Data Independent Acquisition Proteomics Experiments without Spectrum Libraries. *Mol. Cell. Proteom.* **19**, 1088–1103 (2020).
91. Visser, I. & Speekenbrink, M. depmixS4: An R Package for Hidden Markov Models. *J. Stat. Softw.* **36**, 1–21 (2010).
92. Methé, B. A. et al. The psychrophilic lifestyle as revealed by the genome sequence of *Colwellia psychrerythraea* 34H through genomic and proteomic analyses. *Proc. Natl Acad. Sci.* **102**, 10913–10918 (2005).
93. R: The R Project for Statistical Computing. <https://www.r-project.org/>.
94. brooknunn & emmats. Nunn-Lab/Publication-2021-Orcas-Island-Time-Series: Nat.Comm Harmful algal blooms are preceded by a predictable and quantifiable shift in the oceanic microbiome. Zenodo <https://doi.org/10.5281/zenodo.14976385> (2025).

Acknowledgements

We acknowledge the contributions of the following people to the project and thank them for their efforts. The University of Washington Genome Sciences Core Sequencing Facility aided in processing and understanding the metagenomics component of the project. Sam White assisted with the metagenomic database assembly. The Youngren family, specifically Kathy and Jim, for donating their house, boat, and land freely for our use in collecting samples. Mike O'Connell aided with fieldwork. Ardi Kveven and Katherine Dye loaned the YSI Exo1 Sonde probe and contributed to teaching its usage and calibration. Fieldwork support was provided by Kathryn Sharp, Lauren Gonsalves, Sam Hart, Tasman Nunn, Edith Branner, Katherine Riffle, Kusum Dhakar, Angela Currie, Stuart Currie, and Maya Garber-Yonts. We acknowledge that this work was conducted on the lands of the Coast Salish peoples, land that touches the shared waters of all tribes and bands within the Suquamish, Tulalip, and Muckleshoot nations. We acknowledge SoundToxins East Sound Monitor Specialist Lynette Wood for providing relative abundances of phytoplankton during the field season. We thank Mark D. Scheuerell for his guidance on time series statistical analyses. This work was possible due to funding by the NIH to BL Nunn and J Kubanek (NIH NIEHS R21ES034337) and a fellowship to MC Mudge (F31 ES032733-01A1). Fieldwork was funded by a pilot grant from The University of Washington Royal Research Foundation to BL Nunn in 2021 and donations from the Nunn family. Support for personnel was also provided by NSF grants NSF OCE 2401644 to BL Nunn and MD Scheuerell, NSF OCE 2401645 to J Kubanek, NSF OCE 2401646 to TA Rynearson. This work is supported in part by the University of Washington's Proteomics Resource (UWPR95794).

Author contributions

M.C.M., J.K., and B.L.N. designed the study. M.C.M., M.R., G.C., D.L.P., E.T.S., and B.L.N. participated in sample collection. T.A.R. helped with experimental design for efficient sample collection. M.C.M. and E.T.S. performed DNA extractions. M.C.M. processed and analyzed all samples. M.C.M. performed the data analysis, with assistance and feedback on data analysis and interpretation provided by M.R., E.T.S., W.S.N., and B.L.N. M.R. assisted with software development. D.L.P. provided feedback on use of Skyline and mass spectrometry interpretation. J.K. and B.L.N. served as co-investigators on project funding. M.C.M. and B.L.N. wrote the manuscript with input and editing from all authors. All authors reviewed the final manuscript draft.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at <https://doi.org/10.1038/s41467-025-59250-y>.

Correspondence and requests for materials should be addressed to Brook L. Nunn.

Peer review information *Nature Communications* thanks Juliette Smith, and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. A peer review file is available.

Reprints and permissions information is available at <http://www.nature.com/reprints>

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Open Access This article is licensed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License, which permits any non-commercial use, sharing, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if you modified the licensed material. You do not have permission under this licence to share adapted material derived from this article or parts of it. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit <http://creativecommons.org/licenses/by-nc-nd/4.0/>.

© The Author(s) 2025