RESEARCH ARTICLE

Needles in the blue sea: Sub-species specificity in targeted protein biomarker analyses within the vast oceanic microbial metaproteome

Mak A. Saito[1], Alexander Dorsk[1], Anton F. Post[2], Matthew R. McIlvin[1], Michael S. Rappé[3], Giacomo R. DiTullio[4] and Dawn M. Moran[1]

[1] Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution, Woods Hole, MA, USA
[2] Coastal Resources Center, URI Graduate School of Oceanography, Narragansett, RI, USA
[3] Hawaii Institute of Marine Biology, SOEST, University of Hawaii, Kaneohe, HI, USA
[4] Grice Marine Laboratory, College of Charleston, South Carolina, USA

Proteomics has great potential for studies of marine microbial biogeochemistry, yet high microbial diversity in many locales presents us with unique challenges. We addressed this challenge with a targeted metaproteomics workflow for NtcA and P-II, two nitrogen regulatory proteins, and demonstrated its application for cyanobacterial taxa within microbial samples from the Central Pacific Ocean. Using METATRYP, an open-source Python toolkit, we examined the number of shared (redundant) tryptic peptides in representative marine microbes, with the number of tryptic peptides shared between different species typically being 1% or less. The related cyanobacteria *Prochlorococcus* and *Synechococcus* shared an average of $4.8 \pm 1.9\%$ of their tryptic peptides, while shared intraspecies peptides were higher, $13 \pm 15\%$ shared peptides between 12 *Prochlorococcus* genomes. An NtcA peptide was found to target multiple cyanobacteria species, whereas a P-II peptide showed specificity to the high-light *Prochlorococcus* ecotype. Distributions of NtcA and P-II in the Central Pacific Ocean were similar except at the Equator likely due to differential nitrogen stress responses between *Prochlorococcus* and *Synechococcus*. The number of unique tryptic peptides coded for within three combined oceanic microbial metagenomes was estimated to be $\sim 4 \times 10^7$, 1000-fold larger than an individual microbial proteome and 27-fold larger than the human proteome, yet still 20 orders of magnitude lower than the peptide diversity possible in all protein space, implying that peptide mapping algorithms should be able to withstand the added level of complexity in metaproteomic samples.

Received: December 29, 2014
Revised: April 17, 2015
Accepted: June 9, 2015

Keywords:
Biomarkers / Cyanobacteria / Metatryp / Microbiology / MRM / Targeted metaproteomics

Additional supporting information may be found in the online version of this article at the publisher's web-site

1 Introduction

The ocean is an immense environment that creates and maintains habitable conditions on Earth, as well as being of vital economic importance to human society. There are numerous anthropogenic perturbations that impact ocean ecosystems [1], yet there remains considerable uncertainty regarding their long-term effects [2]. Due to the vastness of the oceans and the relatively small number of ocean scientists, it remains a major logistical challenge to characterize the oceans' ecosystems representatively. Typical research efforts involve ship-based expeditions that focus on a particular geographic region for a short period of time (days to weeks). Alternatively coastal (various long-term ecological research sites) and oceanic time-series sampling sites have been initiated to detect long-term

Correspondence: Dr. Mak A. Saito, Marine Chemistry and Geochemistry Department, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, USA
E-mail: msaito@whoi.edu

changes with monthly resolution (e.g., Hawaii and Bermuda Atlantic time series stations). Finally, there are ocean basin "sections" that have surveyed the distribution of key chemical elements and compounds across ocean provinces (e.g., WOCE for nutrients, GEOTRACES for metals). The microbial community of the oceans has been biologically interrogated in various process studies, surface transects, and time-series analyses. These efforts have typically focused on aspects of the microbial productivity and/or diversity present, with characterization of the functional and biochemical capabilities being less common. Arguably, a comprehensive understanding of marine microbial biogeochemistry and its response to ocean change has been logistically and methodologically constrained.

Recent advances in mass spectrometry-based proteomics methodologies offer powerful tools for the analysis of not only single organisms, but also for the study of more complex communities of organisms be they free-living in the natural environment or as microbiomes associated with larger organisms [3, 4]. These communities harbor great biological diversity, containing an abundance of bacterial species as in the case of microbiomes, or very diverse assemblages from all three superkingdoms (Bacteria, Archaea, and Eukaryotes) as in the case of the ocean ecosystem environments. In recent years, several studies have demonstrated the potential to identify proteins and their relative abundances, and most recently, to quantify targeted protein biomarkers in complex natural environments to provide ecosystem and biogeochemical insights [4–6]. Yet, proteomic bioinformatic development has almost exclusively focused on single organisms rather than communities, despite the significant challenge of finding peptide mass "needles" in this ocean size haystack of protein diversity. Furthermore, the reliance of MS-proteomics for the identification of exact matches of tryptic peptides in predicted sequences from genomes can complicate matters. Only a ~10% divergence in the amino acid sequence (90% identity) of a protein can be tolerated before insufficient tryptic peptides remain for identification [7].

Metaproteomics, often defined as the analysis of a complex community of organisms [8], has unique challenges relative to "standard" proteomics of a single organism. By considering these challenges we can evaluate the limitations of existing algorithms and pipelines, as well as provide motivation for future software development. Two primary objectives of metaproteomics at this early stage of study include: (i) the identification and (ii) the quantitation of protein. Protein identification in metaproteomes currently relies on large sequence databases created from both concatenated genomes and metagenome libraries in order to cover as much of the natural diversity of proteins present as possible [3]. Progressing from global discovery proteomics, where maximizing identifications is the primary objective, to targeted (meta)proteomics, now allows identification and quantitation of biomarkers that diagnose the environmental stresses experienced by individual members within the microbial community [5], as well as estimating biogeochemical functions through measurement of key enzymes, and making estimates of potential enzyme activity using specific activity and/or kinetic parameters.

In this manuscript we describe a workflow that combines discovery metaproteomic analyses, genomic in silico analyses of tryptic peptide diversity, and quantitative targeted proteomic measurements on the complex microbial community of the oligotrophic surface ocean. We focus on marine cyanobacteria due to their abundance and importance, in particular the species *Prochlorococcus*, which was discovered in the late 1980s by deploying flow cytometers at sea for the first time. *Prochlorococcus* is now known to be the single largest contributor to carbon fixation on Earth, contributing approximately 10% of photosynthetic activity globally [9, 10]. *Prochlorococcus* lives among other abundant marine microbes, such as the related and also highly abundant cyanobacterium *Synechococcus*, as well as rarer, yet biogeochemically important nitrogen fixing cyanobacteria, such as *Crocosphaera* and *Trichodesmium* and the symbiotic cyanobacterium UCYN-A. *Prochlorococcus* and *Synechococcus* have numerous genomes available based on cultivated isolates from around the world [11,12]. In addition, major alpha and gamma proteobacterial clades such as SAR11 and SAR86, and also recently recognized oceanic Archaea, such as the Thaumarchaeota, are abundant in the oceans. Together these major groups of microbes comprise a majority of free-living marine microbial populations in the open ocean surface layers [13], although this excludes the extensive eukaryotic phytoplankton diversity found predominantly in larger size fractions.

The resulting metaproteomic biomarkers are able to resolve microbial biodiversity to the level of individual marine cyanobacterial species, and in some cases beyond that to specific ecotypes within those species. This specific analysis of functional aspects of marine microbial populations over geographical scales of thousands of kilometers has the potential to be deployed on future oceanographic surveys to detect large-scale changes in the oceans. With major changes known to be occurring throughout the oceanic and coastal ecosystems, this capability to detect ecosystem changes and the nutritional factors that control key biogeochemical processes could be built and deployed to diagnose known and as yet unknown alterations of the oceans.

2 Materials and methods

Oceanic protein samples were collected by high-volume submersible McLane pumps (McLane Research, Falmouth MA) using custom Mini-Mulvfs filtration heads attached to a nonmetallic line on the Research Vessel *Kilo Moana* in 2011. Each sample consisted of ~300 L of seawater filtered through 0.2 Supor membrane filter, with pre-filtration through 3.0 and 51 micron filters, and preserved in RNAlater, which has been shown to effectively preserve cyanobacterial proteins [14], and frozen at –80°C until extraction. One quarter

of the 0.2 micron filters were extracted with an SDS-based protocol, embedded in a tube gel to purify away the detergent and salts [15], alkylated and reduced prior to trypsin digestion as previously described [5] (also see Supporting Information).

Biomarkers for two global nitrogen regulatory proteins were chosen from abundant proteins identified within a metaproteomic discovery dataset generated from samples from this research expedition and location, as previously described [5]. For the purposes of this environmental example, a tryptic peptide from each protein was targeted: the P-II protein (ID-34, VNSVIDAIAEAAK, MW 1299.70) and the NtcA protein (ID-35, LSHQAIAEAIGSTR, MW 1452.76). Comparison of three tryptic peptides from two proteins in this expedition was previously presented and showed good spatial coherence [5], although because of the natural population diversity this practice presents challenges unique to metaproteomics. Absolute quantitation of proteins was conducted using multiple reaction monitoring (MRM) by triple quadrupole mass spectrometry using a Thermo Vantage mass spectrometer and synthetic isotope labeled peptide standards as described previously [16]. Isotopically labeled standards were obtained from JPT Peptide Technologies, which contain a C-terminal peptide tag. The tag was released by tryptic digestion prior to analysis following the manufacturer's protocol. Peptides were chosen with an effort to minimize the presence of methionine and cysteine residues, which can be oxidized and create variability in analyses [17, 18]. Mass spectrometry conditions were optimized for each peptide (collision energy and S-lens), and analyzed using chromatographic scheduling to increase the resolution for each peptide. For P-II, the precursor ion 650.859 (+2) was isolated and fragment ions 1087.5994, 901.4989, and 788.4149 were measured using collision energies of 21, 21, and 23, and S-Lens value of 148 for all. The heavy labeled version of this peptide had a precursor ion of 654.859, and fragment ions 1095.5994, 909.4989, and 796.4149, with identical collision and S-Lens values as the light peptide. For the NtcA protein the precursor ions 485.2634 and 488.5968 for light and heavy peptides were isolated (+3), and fragment ions 604.3414, 533.3042, 420.2201 for the unlabeled peptide and 614.3413, 543.3042, and 430.2201 for the heavy labeled peptide were quantified with an S-Lens value of 77 for all. Peptide abundances were calculated as a peak ratio of the corresponding isotopically labeled internal standard. Each internal standard was examined for its linear performance on the mass spectrometer using standard curves. Chromatographic separation and mass spectrometry were performed using a Paradigm MS4 HPLC (Michrom Bioresources) coupled to a Thermo Vantage TSQ mass spectrometer (Thermo Scientific) via an Advance capillary electrospray source (Michrom Bioresources). Samples were loaded on a peptide CapTrap prior to separation on a Magic C18AQ column (0.2 × 50 mm, 3 μm particle size, 200 Å pore size, Michrom Bioresources). Chromatographic separation was done with a 45 min gradient of 5 to 35% buffer B (where buffer A was 0.1% formic acid in water, Fisher Optima and buffer B was 0.1% formic acid in acetonitrile, Fisher Optima) at 4 μL/min. LOD and LOQ were 0.009 fmol and 0.025 fmol for peptide ID-34 and 0.013 fmol and 0.035 fmol for peptide ID-35, respectively.

A Python software toolkit (METATRYP) was written that ingests microbial genomes and digests them according to proteolytic enzyme rules. The toolkit source code and documentation are available on Github: www.github.com/saitomics/metatryp. The toolkit uses trypsin as the default digestion enzyme but other enzymes can be programmed as regular expressions. The resulting peptides and metadata are stored in a stand alone SQLite database, which can then be queried by command line scripts. Custom user command line SQLite queries are also possible, and examples are provided in toolkit documentation. Ingest and digest parameters can be set to change the digestion enzyme, number of missed cleavages allowed, and minimum and maximum number of amino acids per peptide. The METATRYP toolkit was developed in parallel with the web-server based UNIPEP [19], but has advantages in its use of an offline, portable SQLite database accessible through command-line querying, and in allowing the use of novel microbial genomes and custom sequence files prior to public release that are now common in environmental microbiology fields with the widespread use of inexpensive high-throughput DNA sequencing capabilities. METATRYP installation and use is straightforward with several Bash scripts for data ingestion (digest_and_ingest.sh), SQLite database confirmation (list_taxon_ids.sh), analysis of shared tryptic peptides between genomes (generate_redundancy_tables.sh), and querying of the SQLite database for specific peptide sequences shared between genomes (query_by_sequence.sh). A useful feature of METATRYP is that it allows for sequence variability in sequence queries to identify of homologous sequences that could be targeted (query_by_sequence –max-distance [value]). Redundancy peptide tables can also be easily generated for a subset of genomes within the SQLite database by use of a taxon input file and script parameter (generate_redundancy_tables.sh –taxon-id-file [filename]).

Metaproteome size estimates, as the number of unique tryptic peptides, were made using metagenomic resources, the Chainsaw trypsin digestion program from Proteowizard [20], and custom Bash Shell scripts to count number of peptides of each peptide length and the total number of peptides between 6–22 amino acid length. Genomes and metagenomes were downloaded from NCBI (http://www.ncbi.nlm.nih.gov/), the Joint Genome Institute Integrated Microbial Genome Portal (http://img.jgi.doe.gov/), and at iMicrobe at http://data.imicrobe.us/). The combined Pacific metaproteome was created using metagenomic datasets from Station Aloha, Line P, and Saanitch Inlet. Metagenomic samples were typically filtered onto 0.2 micron filters to concentrate prokaryotic cells and prefiltered with glass fiber filters to remove larger eukaryotic cells [21]. Metaproteome sequences were translated from raw sequence or partial genomic reads,

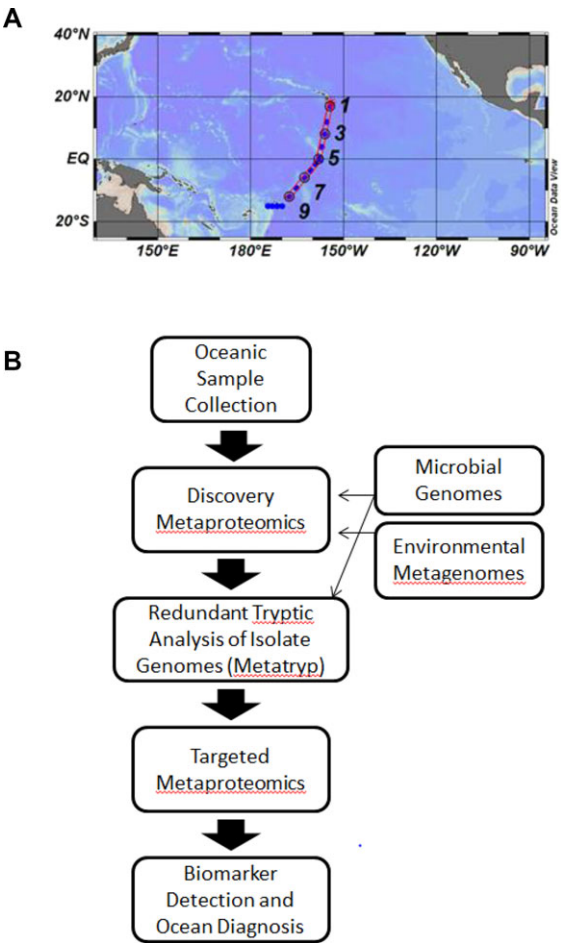

Figure 1. (A) Sampling region for the METZYME expedition on the R/V *Kilo Moana* in 2011. (B) Targeted metaproteomic workflow for quantitation of species-specific oceanic biomarkers.

and hence represent tryptic, semi-tryptic, or truncated sequences (for sequences at the ends of DNA reads).

3 Results

We conducted targeted metaproteomic analyses of a Pacific Ocean microbial community using the workflow shown in Fig. 1. Specifically, abundant peptides associated with proteins of interest from field metaproteomes were selected and subjected to in silico analysis of the occurrence of those peptides within representative microbial isolate genomes (Fig. 2). Next targeted metaproteomic assays were designed, optimized and applied to the original samples (Fig. 3). The discovery and targeted mass spectrometry-based proteomics methodologies and their environmental implications for nutrient stress in Central Pacific *Prochlorococcus* were previously described [5], and here we specifically elaborate on the challenges associated with targeting individual species in complex communities using in silico analysis and its technical implications. Specific examples of inter- and intra-species level specificity are described along with their implications for targeted analyses in the marine environment.

Key to the practical implementation of targeted proteomics in natural communities with complex assemblages of microbes is an ability to assess the taxa associated with each targeted peptide. In essence, if one is designing a targeted protein assay for a protein it is important to know how many species may contain the peptide of choice in order to correctly assign its microbial taxonomic origin. The Python METATRYP library we developed creates a SQL database compatible output that can be easily searched to identify occurrences of peptides in representative microbial proteomes, translated in silico from genomes. The number and percentage of tryptic peptides shared in common between pairs of ~50 selected marine microbial genomes representative of the pelagic oligotrophic ("blue water") ocean were calculated to assess a broad sense of the specific metaproteomic capabilities (Table 1 and Fig. 2). Because some genomes are larger than others (as shown in the varying total number of tryptic peptides per genome in Table 1), the *direction* of the pairwise comparison influences the calculated percentage of shared peptides. The heatmap and associated values in Fig. 2 allow this bidirectional pairwise comparison where genomes on the X-axis refer to the genome (and the total number of tryptic peptides coded within) being compared to (e.g., the denominator genome). The number of peptides shared *within* a species can range from ~5% to as high as ~50–55% for closely-related strains (e.g., *Synechococcus* BL107 and *Synechococcus* PCC9902; *Prochlorococcus* MIT9601 and *Prochlorococcus* MIT9301). Analysis of 12 *Prochlorococcus* genomes in this manner identified 13 ± 15% shared peptides between them, reflective of the broad range within the *Prochlorococcus* species. Comparisons *between* closely related species, such as the marine cyanobacteria *Prochlorococcus* and *Synechococcus* tend to range between 1 and 10%, with an average of 4.8 ± 1.9 for *Prochlorococcus* peptides within *Synechococcus* genomes ($n = 10$; marine strains only). More distantly related nitrogen fixing cyanobacteria species such as *Trichodesmium* and UCYN-A have 2% or lower shared tryptic peptides within *Prochlorococcus* and marine *Synechococcus*, as well as being easily physically separated by being in a larger filtration size fraction. Similarly, other bacterial species such as the highly abundant *Pelagibacter* SAR11 group have 0.3% or fewer shared peptides in common with the marine cyanobacteria. Two non-marine microbes (*E.coli* GCA and *Pseudomonas aeruginosa* PA7) were also included as controls and had ~0.5% tryptic peptides shared in common with marine cyanobacteria *Prochlorococcus* and *Synechococcus*. Together these analyses demonstrate that most tryptic peptides within the microbial species examined here are available for species-level targeting. Given that *Prochlorococcus*, *Synechococcus*, and the SAR11 clade are considered to be three of the most abundant free-living microbes in the subtropical and tropical open ocean environments, this low level of shared tryptic peptides is

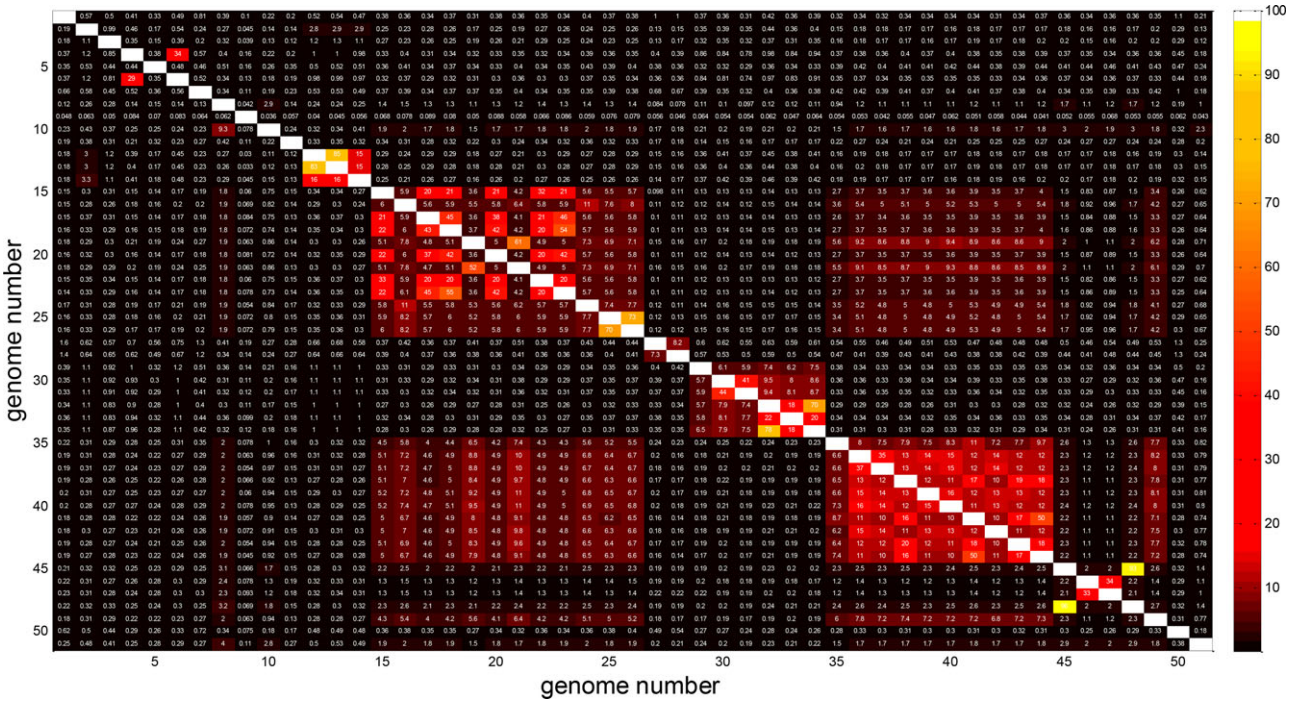

Figure 2. Heatmap of the percent tryptic peptides shared between 51 microbial genomes (see Table 1 for genome number key). Each pairwise comparison is represented by a square with the value of the percent shared tryptics embedded. Each genome's comparison with itself is represented by the white diagonal line. *Prochlorococcus* are found between genomes 15 and 26, and *Synechococcus* between 35 and 49. Interspecies comparisons of *Prochlorococcus* to *Synechococcus* genomes are found off the diagonal between genomes 15 and 26 on the vertical axis and 35 and 49 on the horizontal axis, or vice versa for *Synechococcus* to *Prochlorococcus* comparisons. A black and white version of this heatmap is available in the Supporting Information.

particularly encouraging for deployment of targeted metaproteomic methods.

To demonstrate the issues involved with unique and shared peptides in marine microbes when deploying targeted metaproteomic assays, we selected two example tryptic peptides from two global nitrogen regulatory proteins, P-II and NtcA. Specific peptides corresponding to each protein that were abundant in the global discovery dataset were selected and examined using METATRYP queries, where their presence within 51 microbial genomes was characterized (Table 1). These two nitrogen stress proteins provide an interesting example case, where their vertical distributions are generally similar across the Central Pacific Ocean, increasing in abundance towards the surface and consistent with nitrogen stress in the North Pacific Subtropical Gyre (Fig. 3), prior to onset of iron limitation at the Equator [5]. Yet these two peptides showed differences in species specificity by METATRYP analysis: the NtcA tryptic peptide sequence was found in high and low light ecotypes of the *Prochlorococcus* species, in all of the marine *Synechococcus*, and the three nitrogen fixing cyanobacteria representatives (*Crocosphaera*, *Trichodesmium*, and UCYN-A). In contrast, the P-II targeted peptide was exclusively found in the high-light ecotype of *Prochlorococcus* (Table 1). While both of these nitrogen regulatory proteins, as represented by their targeted peptides, clearly showed nitrogen stress when nitrate is scarcer in the North Pacific Subtropical Gyre (NPSG; Fig. 3A–E), they also showed two subtle differences: first, the NtcA protein persisted in the transition waters between NPSG and the equatorial Pacific at stations 3 and 5 (Fig. 3), but P-II had declined significantly or disappeared by those stations, respectively. Second, NtcA showed higher abundances at the shallowest depth compared to P-II at stations 1 and 3. Comparison of biomarker:biomarker distributions (Fig. 3F) illustrates both of these trends, where station 5 has a steep slope indicating low abundances of P-II relative to NtcA there, and the shallowest depth at station 5 (the last point on the line) jogs back towards the left, consistent with a decrease in P-II relative to NtcA.

4 Discussion

While both P-II and NtcA biomarkers showed nitrogen stress in the photic zone of the North Pacific Subtropical Gyre and were consistent with similar distributions of a urea transporter, there were subtle differences in their distributions, particularly at the Equatorial Station 5 (Fig. 3). These results illustrate the value of the METATRYP analysis, allowing a teasing apart of potential taxonomic interferences. Based on

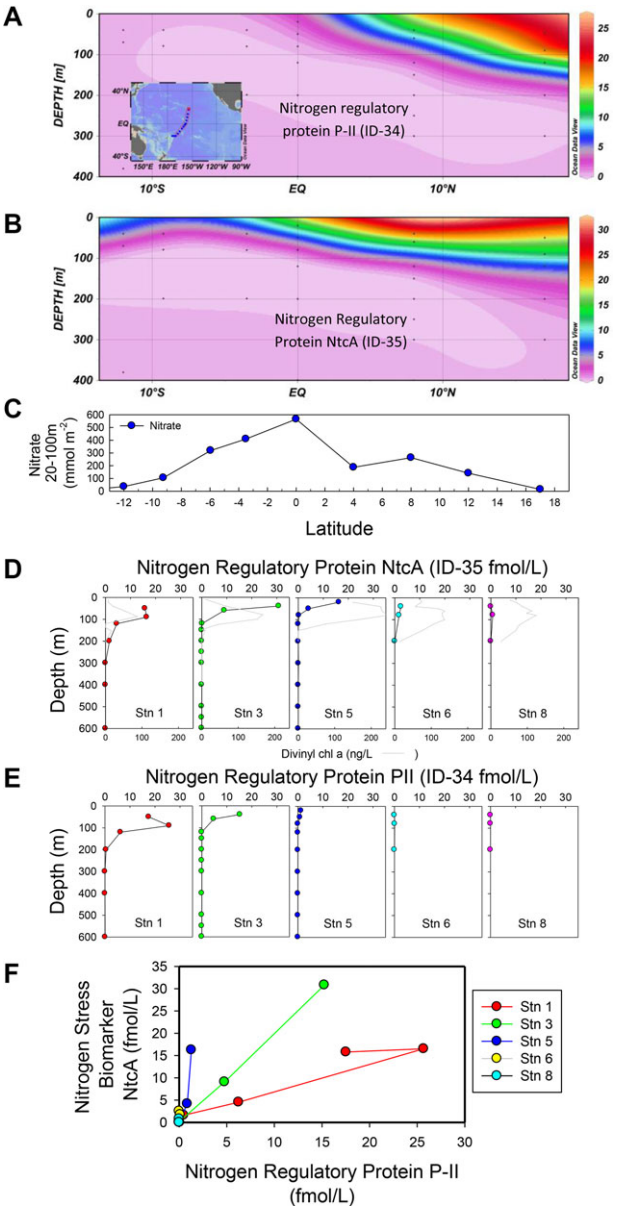

Figure 3. Comparison of the oceanic water column distributions of cyanobacterial nitrogen regulatory proteins P-II (A, E) and NtcA (B, E) in the Central Pacific Ocean in vertical profiles and as an ocean section. (C) Integrated photic zone nitrate concentrations (20–100 m). Zero values measured below the sunlit photic zone, consistent with the measured distribution of these photosynthetic microbes as measured by the unique *Prochlorococcus* pigment divinyl chlorophyll *a* distributions in panel D. (F) Comparison of NtcA versus P-II showed cohesive responses within each station, but varied across stations, likely indicative of different nitrogen stress levels in *Prochlorococcus* and *Synechococcus*. For example, at Station 1 in the Gyre both microbes were likely nitrogen stressed where nitrogen is scarcest, compared to Station 5 on the equator where nitrogen availability was higher and high-light ecotype *Prochlorococcus* is no longer N stressed but displayed iron stress as described in Saito et al. 2014 [5]. Lines in panel F follow trends of decreasing depth, with the deepest samples at the origin where cyanobacterial biomass was least abundant.

the analysis shown in Table 1, we can conclude P-II peptide was specific to *Prochlorococcus*, with the caveat that this interpretation is based on the genomes utilized in the analysis. Yet by the same analysis the NtcA peptide sequence was also found in other cyanobacteria, in particular the abundant *Synechococcus* which often co-occurs with *Prochlorococcus*. Other cyanobacterial populations also contained the targeted NtcA peptide such as *Trichodesmium* and *Crocosphaera*, yet both species are less abundant and would likely have been removed by the size fractionated filtration (0.2–3 micron). As a result, a plausible explanation for the persistence of NtcA at the Equatorial region is that *Synechococcus* continued to experience nitrogen stress at the elevated nitrogen abundances found with Equatorial Upwelling while *Prochlorococcus* ceased to, and that both microbial species could have contributed to this biomarker's distribution across this section. This interpretation is consistent with the larger cell size of *Synechococus* relative to *Prochlorococcus* and the associated advantages for nutrient acquisition that come with the smaller surface-area to volume ratio [22]. As a result, the P-II protein may be a more specific diagnostic of nitrogen stress for *Prochlorococcus*, in addition to the *Prochlorococcus* urea transporter described previously [5], given their species and ecotype-level resolving power (Table 1), while maintaining the diagnostic principle as both P-II and UrtA expression is controlled by NtcA. Although further studies confirming this hypothesized phenomenon would be useful, this example demonstrates how the METATRYP analysis allows specific interpretations of the taxa potentially being measured by each biomarker.

These observations also demonstrate the added value of global discovery-driven metaproteomic data in discovering novel biomarkers for use in ecosystem diagnosis. While NtcA had been previously characterized as a potential nitrogen stress biomarker in the marine cyanobacterium *Synechococcus* [23, 24], less is known about the response of the P-II protein, particularly in the marine cyanobacteria [25–27]. This field examination along a natural gradient showed P-II to be a strong candidate for a biomarker of nitrogen stress. Laboratory studies on non-marine cyanobacteria (*Synechococcus* PCC6803) have found P-II to have distinct nitrogen regulatory functions, and it plays an important role at the intersection of carbon and nitrogen metabolism [26, 28]. In a transcriptome study of two *Prochlorococcus* strains, mixed responses of P-II protein were observed during short-term nitrogen deprivation [25]. These differences may be strain specific, where the MIT9313 strain that lacked a P-II response is a low-light ecotype (Table 1) found near the bottom of the photic zone where dissolved inorganic nitrogen is more abundant, whereas the MED4 strain (also known as CCMP1986) studied is a high-light ecotype strain that lives in nitrogen depleted waters [12]. Similarly, no obvious response by antibody-western blotting method was observed in a 24 h nitrogen deprivation experiment in *Prochlorococcus* strain PCC9511 [27]. In contrast to these laboratory studies, this Central Pacific dataset showed a coherence between P-II and NtcA responses. Interestingly, the specific nitrogen regulatory P-II peptide discovered and

Table 1. Marine microbial genomes used for redundant (shared) tryptic peptide analyses and the total number of tryptic peptides within each

Genome number	Taxon	Phylum/Class/Ecotype	Tryptic peptides	P-II (ID-34)	NtcA (ID-35)
1	*E. coli GCA*	Gammaproteobacteria; non-marine	71413	–	–
2	*HIMB114*	Alphaproteobacteria	23891	–	–
3	*HIMB59*	Alphaproteobacteria	26119	–	–
4	*Nitrobacter 311*	Alphaproteobacteria	63263	–	–
5	*Nitrobacter defluvii*	Alphaproteobacteria	74503	–	–
6	*Nitrobacter winogradski*	Alphaproteobacteria	53999	–	–
7	*Nitrobacter Nb–211*	Alphaproteobacteria	57933	–	–
8	*Atelocyanobacterium thalassa (UCYN–A)*	Cyanobacteria	22612	–	Exact
9	*Cenarchaeum symbiosumA*	Thaumarchaeota	33399	–	–
10	*Crocosphaera 8501*	Cyanobacteria	72663	–	Exact
11	*Kuenenia stuttgartiensis*	Planctomycetes	67145	–	–
12	*Peligibacter 1002*	Alphaproteobacteria	25009	–	–
13	*Peligibacter 1062*	Alphaproteobacteria	24647	–	–
14	*Pelagibacter 7211*	Alphaproteobacteria	26596	–	–
15	*Prochlorococcus CCMP1986 (MED4)*	Cyanobacteria; *High–light Ecotype*	29305	-1	Exact
16	*Prochlorococcus MIT9211*	Cyanobacteria; *Low–light Ecotype*	29584	–	Exact
17	*Prochlorococcus MIT9215*	Cyanobacteria; *High–light Ecotype*	31352	Exact	Exact
18	*Prochlorococcus MIT9301*	Cyanobacteria; *High–light Ecotype*	29937	Exact	Exact
19	*Prochlorococcus MIT9303*	Cyanobacteria; *Low–light Ecotype*	41781	-4	-1
20	*Prochlorococcus MIT9312*	Cyanobacteria; *High–light Ecotype*	30535	Exact	Exact
21	*Prochlorococcus MIT9313*	Cyanobacteria; *Low–light Ecotype*	36064	-4	-1
22	*Prochlorococcus MIT9515*	Cyanobacteria; *High–light Ecotype*	30234	-1	Exact
23	*Prochlorococcus AS9601*	Cyanobacteria; *High–light Ecotype*	30324	Exact	Exact
24	*Prochlorococcus 1375 (SS120)*	Cyanobacteria; *Low–light Ecotype*	30441	–	-1
25	*Prochlorococcus NATL1A*	Cyanobacteria; *Low–light Ecotype*	31679	–	Exact
26	*Prochlorococcus NATL2a*	Cyanobacteria; *Low–light Ecotype*	30457	–	Exact
27	*Pseudomonas PA7*	Gammaproteobacteria; non–marine	111368	–	–
28	*Pseudomonas putida*	Gammaproteobacteria; non–marine	100112	–	–
29	*Roseobacter sp. MED193*	Alphaproteobacteria	74759	–	–
30	*Roseobacter denitrificans OCh 114*	Alphaproteobacteria	69803	–	–
31	*Roseobacter litoralis Och 149*	Alphaproteobacteria	74525	–	–
32	*Sulfitobacter sp. EE–36*	Alphaproteobacteria	58154	–	–
33	*Sulfitobacter sp. GAI–101*	Alphaproteobacteria	70434	–	–
34	*Sulfitobacter sp. NAS–14*	Alphaproteobacteria	64122	–	–
35	*Synechococcus WH5701*	Cyanobacteria	48405	–	Exact
36	*Synechococcus WH7803*	Cyanobacteria	39836	-4	Exact
37	*Synechococcus WH7805*	Cyanobacteria	42586	-4	Exact
38	*Synechococcus WH8102*	Cyanobacteria	39990	–	Exact
39	*Synechococcus RS9916*	Cyanobacteria	42530	–	Exact
40	*Synechococcus RS9917*	Cyanobacteria	42160	-4	-1
41	*Synechococcus BL107*	Cyanobacteria	37482	–	Exact
42	*Synechococcus CC9311*	Cyanobacteria	41294	-3	Exact
43	*Synechococcus CC9605*	Cyanobacteria	40526	–	Exact
44	*Synechococcus CC9902*	Cyanobacteria	36932	–	Exact
45	*Synechocystis PCC6803*	Cyanobacteria; non–marine	42272	–	Exact
46	*Synechcooccus JA–2–3Ba*	Cyanobacteria; non–marine (Yellowstone)	46875	–	Exact
47	*Synechococcus JA–3–3Ab*	Cyanobacteria; non–marine (Yellowstone)	45297	–	Exact
48	*Synechococcus PCC7942*	Cyanobacteria	43521	–	Exact
49	*Synechococcus RCC307*	Cyanobacteria	37847	–	Exact
50	*Thiomicrospira crunogena*	Gammaproteobacteria; hydrothermal vent	40564	–	–
51	*Trichodesmium sp. IMS101*	Cyanobacteria; diazotroph	88483	–	Exact

Genome numbers listed correspond to the pairwise analyses in Fig. 2. Tryptic peptides representing the global nitrogen regulatory proteins P-II and NtcA, with exact matches or one to four amino acid variants shown (as negative numbers), and dashes for not present. Amino acid variants were not detected by this targeted method, but could be added in future analyses. ID-34 and ID-35 refers to the unique peptide identification number used in Saito et al., 2014.

targeted in our study was not present as an exact match in either genome of the strains studied in the Tolonen study (MED4 and MIT9313; Table 1, present with one and four amino acid variants from the targeted sequence), likely due to the biogeographical differences as we expect the high-light II ecotype (HL II) of *Prochlorococcus* to be abundant in the Central Pacific Ocean. In more temperate climate zones where the high-light I ecotype (HL I; including MED4) of *Prochlorococcus* is found [9, 29], alternate peptide biomarkers could be employed to detect P-II. In the case of these biomarkers we are fortunate to have multiple signals indicating and confirming the diagnosis of ecosystem level nitrogen scarcity (P-II, NtcA, and urea transporters) for the Central Pacific Ocean ecosystem [5].

For the abundant marine cyanobacteria *Prochlorococcus*, we have demonstrated the capability of the development and deployment of species and ecotype level specific biomarkers for nitrogen and other nutrient stresses within natural ecosystems. Numerous genomes are available for the major marine cyanobacteria *Prochlorococcus* and *Synechococcus*, with most of them as closed (complete) genomes, thus offering the level of resolution as depicted here. Yet other marine microbial taxa that are extremely abundant in the oceans but even more difficult to cultivate such as SAR11, SAR86, and the Thaumarcheota, have fewer genomes from cultivated isolates available. For example, there are only three closed genomes (a requirement for accurate tryptic digest prediction) available for the marine pelagic Thaumarcheota and SAR86 at present [30, 31], and hence identifying tryptic peptide biomarkers will be more challenging relative to the cyanobacteria until more sequence information becomes available. Similarly, distant members of the *Peligibacter* SAR11 clade such as tropical coastal HIMB59 and HIMB114 [32] have only ~3% shared tryptic peptides with temperate coastal ocean strains *Peligibacter* 1002 and 1062 and the subtropical Atlantic Ocean strain 7211 (Fig. 2), implying there is considerable heterotrophic bacterial diversity that remains to be identified and incorporated into metaproteomic interpretations. Careful cultivation efforts and single cell genome sequencing will provide assistance in targeting these underrepresented taxa, although the latter method does not currently produce closed genomes, thus hindering the ability to fully document a peptide's shared use. Future applications could use curated metagenome assemblies to further infer taxon associations of tryptic peptides.

In addition, because this targeted metaproteomic workflow (Fig. 1) relies on bottom-up (shotgun) datasets for peptide targets, there are parallel issues associated with the "assembly" of identified peptides to corresponding proteins in the natural environment. Because sequence diversity exists within strains and species of metaproteomic datasets, as described above, it can be difficult to be entirely confident of the assignment of tryptic peptides to protein sequences since there is the possibility that there are multiple related proteins present within a sample that share tryptic peptide sequences. Utilizing metagenome assemblies, ideally from similar geographical and temporal environmental space, could enhance our protein sequence assembly capability. One implication of this is that the popular proteomics practice of mapping multiple unique tryptic peptides to a single protein in model organisms may be difficult to adhere to in targeted metaproteomics studies, at least in the short-term, since the sequence diversity and resultant tryptic peptides shift subtly across ecotypes/strains and thus potentially across geographic regions. Consequently, a useful first approach to quantitating proteins within complex environmental communities is to focus on each tryptic peptide (and multiple tryptic peptides as in this study) as the fundamental unit of analysis and quantitation, and as a representation of the targeted protein. Future studies could aim to quantitatively measure the abundance of a protein "population" by measuring multiple closely related peptides to capture the diversity associated within specific proteins of a species or at the sub-species level within a population (e.g., ecotypes). Similar peptides for those targeted here are present within other cyanobacterial genomes, as shown in Table 1 (listed as –1 to –4). A capability for analysis of the protein population would require a comprehensive assessment of the microbial diversity present, fortunately a considerable capability for this has been already being established in the open ocean surface environment using metagenomic sequencing [33].

Another challenge associated with the discovery of biomarkers for ecosystem diagnosis concerns the ability to conduct discovery-level proteomics of complex mixed community assemblages. In short, if biomarker peptides cannot be discovered from the complex metaproteome spectral dataset, then the development of targeted assays is also hindered. The present approach for discovery metaproteomics has utilized peptide mapping algorithms designed for single organism analysis (e.g., SEQUEST, X!Tandem; [34, 35]), rather than for the large amounts of translated DNA and the numerous redundant protein sequences that result from utilizing a compilation of many genomes and metagenomic datasets. This approach of using large and continually growing genome sequence libraries remains somewhat unsatisfying as a long-term approach to metaproteomic research. Natural environments are subject to evolutionary modifications of amino acid sequences, as well as shifts in abundances of co-occurring species that could increase the prevalence of rare microbes to dominating ones. Being continually tied to reanalysis of genome and metagenome samples seems an inefficient means to capture small variations in amino acid sequences, particularly if large geographic regions such as the oceans are intended to be studied. In short, do we need to continually resequence the DNA of the oceans' microbial diversity in order to maintain a metaproteomic diagnostic capability? Alternatives to this approach could include de novo proteomic sequencing algorithms (algorithms that do not require genomic databases for the identification of proteins within mass spectra) or the incorporation of point mutations into peptide mapping algorithms.

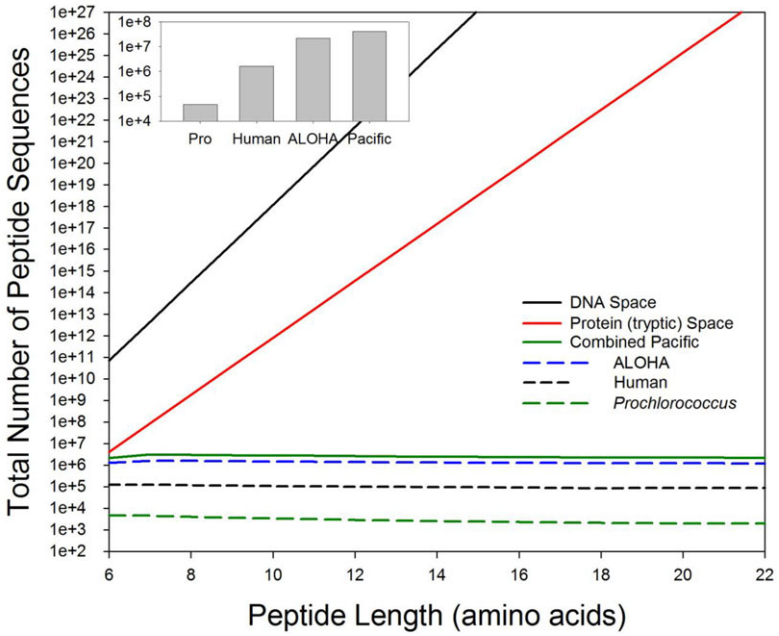

Figure 4. Estimates of the number of unique tryptic peptides found in samples relevant to ocean studies. The number of unique tryptic peptides within (1) the *Prochlorococcus* MED4 (CCMP1986), (2) the human proteome (for comparison), (3) the Station ALOHA (near Hawaii) metaproteome using translated genomic sequences [21], (4) a combined Pacific metaproteome database, (5) all possible protein space (using 21 possible amino acids and no post-translational modifications), and (6) all possible DNA space - untranslated, unique DNA sequences for 18–66 base pairs in length. Observed metaproteomes and proteome peptide diversity decreases with increasing tryptic peptide length, likely due to the occurrence of tryptic cleavage sites, in contrast to the much greater number of possible peptide sequences in protein space. Inset: cumulative number of unique tryptic peptides 6–22 amino acids in length for *Prochlorococcus* (Pro), human, the Pacific ALOHA station metaproteome, and the combined Pacific microbial metaproteome. These estimates for metaproteome space are based on the extent of genome sequencing depth, where deeper sequencing could include rarer microbes.

Given that environmental protein identification has essentially relied on a repurposing of peptide mapping algorithms designed for simpler *single* organism proteomics, it is useful to quantitatively estimate how much more difficult the task of protein identification is for metaproteomics of complex microbial ecosystems. A simple way to approach this is to estimate the number of tryptic peptides that are theoretically possible for mass spectrometry amenable amino acid space (e.g., ~6–22 length peptides; excluding post-translational modifications) and compare this to the sum of unique tryptic peptides coded in microbial proteomes and metaproteomes using translated DNA sequences. While the former value is not realistic relative to estimates that a limited number of protein folds are likely utilized by life (<10 000) [36], this potential diversity is still reflective of the current worst-case computational problem of identifying all protein sequences within environmental samples. In this manner, we can estimate the extent of amino acid space that is being utilized by microbial genomes and metaproteomes in microbial populations of the oceans, compared to what is theoretically possible. To estimate all possible amino acid space for peptides, we focused on bottom-up amenable tryptic peptides of length 6–22 amino acids, where the abundance of all possible amino acid combinations is equal to:

$$\sum_{S=6}^{22} \left(21^{S-1} + 2\right) \tag{1}$$

where S is the sequence length, and the carboxy terminus is required to be one of the two tryptic compatible residues (lysine or arginine). We are including selenocysteine as an alternative amino acid (as a 21st possible amino acid) based on its use in certain marine algae. We ignored the interference of proline when adjacent to the tryptic cleavage site, and missed tryptic cleavage sites, for simplicity here, which would have minor opposing influences on the total amount of peptide diversity. For comparison, the corresponding DNA-based information space for peptides was simply calculated as:

$$\sum_{S=6}^{22} 4^{(S\times 3)} \tag{2}$$

where the peptide length (S) is tripled to convert to DNA sequence and the potential nucleotide combinations. Alternatively, this could also be calculated for codon usage and other restrictions to yield significantly lower numbers of potential DNA sequences.

The sum of all possible sequence variants on the DNA and amino acid level for peptides of length 6 to 22 was calculated to be 6.1e27 and 5.5e39 for amino acid and DNA space, respectively. For comparison, the total number of unique tryptic peptides of the same length range coded for within the genome of the cyanobacterium *Prochlorococcus* was 47197 (strain MED4/CCMP1986), compared to 1646039 within the human proteome, as coded for in the genome (Fig. 4 inset). The estimated utilized metaproteome space at Station ALOHA metagenome dataset near the Hawaiian Islands had 2.2e7 unique tryptic peptides, and a combined Pacific Ocean microbial metaproteome made of three metagenomic projects (see Section 2) had 4.2e7 unique tryptic peptides. Note that this is restricted to the microbial size fraction, using 0.2 micron filters and prefilters to remove most eukaryotic cells, as well as being limited to the sequencing depth of those datasets. Increased sequencing depth would add rarer microbial sequences. This current oceanic metaproteome estimate is ~20 orders of magnitude smaller than the theoretical max-

imal possible number of tryptic peptides for the size range of peptide sequence length. An important characteristic of the real peptide diversity is that the number of tryptic peptides per amino acid length *decreases* as the sequence length becomes longer even in the complex metaproteome samples (at >8 amino acid peptide length), in contrast to the theoretical peptide diversity that increases exponentially with each added amino acid (Fig. 4). This difference is perhaps due to both structural limitations as well as the occurrence for tryptic cleavage sites that prevent the accumulation of longer peptide diversity. The fact that 20-fold less amino acid space is utilized, compared to what is theoretically possible, implies that there are significant design constraints placed on the protein sequence, including from specific protein folds and their secondary and tertiary structural requirements, as well as to the coordination environments for metal ions that can arise from unique positional dependencies (e.g., zinc finger motifs). As a result, much of amino acid space may either not be useable and/or has not been explored by life as of yet [33, 36]. The implication of this calculation is that the large expanse of potential amino acid space does not necessarily need to be searched *de novo* to capture the existing metaproteome diversity in nature, but rather algorithms that can increase the diversity search space by only several orders more than available metagenomic resources could capture much of the microbial diversity. Moreover, because this estimate of metaproteome extent is estimated by use of translated DNA sequences, it implies that the workflow utilized here (Fig. 1) coupled to high-quality metagenomic resources is a good first approach for environmental metaproteomic analysis and ecosystem diagnosis, and that incremental improvements in algorithm and sequence database use could improve these efforts as opposed to the development of new de novo sequencing approaches.

It is interesting to contrast the limited extent of proteome space with that of chemical molecule space as well, which is estimated to be on the order of 10^{60} [37], and is roughly 38 orders of magnitude greater than the calculated peptide space described above and ~53 orders of magnitude greater than our observed utilized metaproteome space. While the extent of chemical diversity space used in the natural environment is difficult to know, it is thought that much of this space may be populated in such low quantities as to prevent active utilization or destruction of those molecules. In contrast, the polymeric nature of proteins and peptides and their biological production by processes constrained by natural selection and functional requirements appears to limit the use of possible peptide space considerably relative to all possible amino acid space.

Together, the successful demonstration of targeted analyses within the highly complex metaproteome environment of the oceans [5], as well as in silico analyses of redundant peptides and limited use of potential peptide sequence space, imply that the deployment of targeted metaproteomic analyses into the vast oceans for ecosystem and biogeochemical diagnosis is a feasible enterprise. While the development of mass spectrometry proteomic technology has been motivated by biomedical needs, the impressive emerging capabilities appear to be of potentially great use to the smaller community of scientists involved in studying and diagnosing the largest ecosystem on Earth. Given the unprecedented rapid and global scale of changes in this ecosystem [1], a proteomic-based diagnostic system could be a valuable tool towards developing sustainable human economies.

We thank the Captain and Crew of the R/V Kilo Moana, and colleagues Carl Lamborg, Alyson Santoro, Tyler Goepfert, Nicholas Hawco, David Wang, and Prentiss Balcolm for technical and sampling assistance during the METZYME expedition in 2011, and Chris Dupont for comments on the manuscript. This research was funded by the Gordon and Betty Moore Foundation and the US National Science Foundation under grant numbers 3782, 3934, OCE-1260233, OCE-1233261, OCE-1220484, OCE-1333212 and OCE-1155566, and the Center for Microbial Oceanography Research and Education (C-MORE).

The authors have declared no conflict of interest.

5 References

[1] Doney, S. C., The growing human footprint on coastal and open-ocean biogeochemistry. *Science* 2010, *328*, 1512–1516.

[2] Boyd, P. W., Strzepek, R., Fu, F., Hutchins, D. A., Environmental control of open-ocean phytoplankton groups: Now and in the future. *Limnol. Oceanogr.* 2010, *55*, 1353–1376.

[3] Morris, R. M., Nunn, B. L., Frazar, C., Goodlett, D. R. et al., Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J.* 2010, *4*, 673–685.

[4] VerBerkmoes, N. C., Denef, V. J., Hettich, R. L., Banfield, J. F., Systems biology: functional analysis of natural microbial consortia using community proteomics. *Nat. Rev. Microbiol.* 2009, *7*, 196–205.

[5] Saito, M. A., McIlvin, M. R., Moran, D. M., Goepfert, T. J. et al., Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. *Science* 2014, *345*, 1173–1177.

[6] Bertrand, E. M., Moran, D. M., McIlvin, M. R., Hoffman, J. M. et al., Methionine synthase interreplacement in diatom cultures and communities: Implications for the persistence of B_{12} use by eukaryotic phytoplankton. *Limnol. Oceanogr.* 2013, *58*, 1431–1450. doi: 10.4319/lo.2013.1458.1434.1431.

[7] Denef, V. J., Shah, M. B., VerBerkmoes, N. C., Hettich, R. L., Banfield, J. F., Implications of strain- and species-level sequence divergence for community and isolate shotgun proteomic analysis. *J. Proteome Res.* 2007, *6*, 3152–3161.

[8] Sowell, S. M., Wilhelm, L. J., Norbeck, A. D., Lipton, M. S. et al., Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. *ISME J.* 2008, *3*, 1–13.

[9] Zwirglmaier, K., Jardillier, L., Ostrowski, M., Mazard, S. et al., Global phylogeography of marine *Synechococcus* and *Prochlorococcus* reveals a distinct partitioning of lineages

among oceanic biomes. *Environ. Microbiol.* 2008, *10*, 147–161.

[10] Flombaum, P., Gallegos, J. L., Gordillo, R. A., Rincón, J. et al., Present and future global distributions of the marine Cyanobacteria *Prochlorococcus* and *Synechococcus*. *Proc. Natl. Acad. Sci.* 2013, *110*, 9824–9829.

[11] Dufresne, A., Ostrowski, M., Scanlan, D. J., Garczarek, L. et al., Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. *Genome Biol.* 2008, *9*, R90.

[12] Rocap, G., Larimer, F. W., Lamerdin, J., Malfatti, S. et al., Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. *Nature* 2003, *424*, 1042–1047.

[13] Yooseph, S., Nealson, K. H., Rusch, D. B., McCrow, J. P. et al., Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature* 2010, *468*, 60–66.

[14] Saito, M. A., Bulygin, V. V., Moran, D. M., Taylor, C., Scholin, C., Examination of microbial proteome preservation techniques applicable to autonomous environmental sample collection. *Front. Microbiol.* 2011, *2*, 215.

[15] Lu, X., Zhu, H., Tube-gel digestion. *Mol. Cell. Proteomics* 2006, *M500138-MCP200*, 1948–1958.

[16] Saito, M. A., Bertrand, E. M., Dutkiewicz, S., Bulygin, V. V. et al., Iron conservation by reduction of metalloenzyme inventories in the marine diazotroph *Crocosphaera watsonii*. *Proc. Natl. Acad. Sci.* 2011, *108*, 2184–2189.

[17] Lange, V., Malmstrom, J. A., Didion, J., King, N. L. et al., Targeted quantitative analysis of *Streptococcus pyogenes* virulence factors by multiple reaction monitoring. *Mol. Cell. Proteomics* 2008, *7*, 1489–1500.

[18] Stahl-Zeng, J., Lange, V., Ossola, R., Eckhardt, K. et al., High sensitivity detection of plasma proteins by multiple reaction monitoring of N-glycosites. Mol. Cell. Proteomics 2007, *6*, 1809–1817.

[19] Mesuere, B., Devreese, B., Debyser, G., Aerts, M. et al., Unipept: tryptic peptide-based biodiversity analysis of metaproteome samples. *J. Proteome Res.* 2012, *11*, 5773–5780.

[20] Chambers, M. C., Maclean, B., Burke, R., Amodei, D. et al., A cross-platform toolkit for mass spectrometry and proteomics. *Nat. Biotechnol.* 2012, *30*, 918–920.

[21] DeLong, E. F., Preston, C. M., Mincer, T., Rich, V. et al., Community genomics among stratified microbial assemblages in the ocean's interior. *Science* 2006, *311*, 496–503.

[22] Chisholm, S. W., *Primary productivity and biogeochemical cycles in the sea*, Springer, New York 1992, pp. 213–237.

[23] Lindell, D.P. S., Al Qutob, M., David, E., Korpal, T. et al., Post expression of the N-stress response gene ntcA reveals N-sufficient *Synechococcus* populations in the oligotrophic northern Red Sea. *Limnol. Oceanogr.* 2005, *50*, 1932–1944.

[24] Lindell, D., Post, A. F., Ecological aspects of ntcA gene expression and its use as an indicator of the nitrogen status of marine *Synechococcus* spp. *Appl. Environ. Microbiol.* 2001, *67*, 3340–3349.

[25] Tolonen, A. C., Aach, J., Lindell, D., Johnson, Z. I. et al., Global gene expression of *Prochlorococcus* ecotypes in response to changes in nitrogen availability. *Mol. Syst. Biol.*2006, *2*.

[26] Forchhammer, K., Global carbon/nitrogen control by PII signal transduction in cyanobacteria: from signals to targets. *FEMS Microbiol. Rev.* 2004, *28*, 319–333.

[27] Palinska, K. A., Laloui, W., Bédu, S., Loiseaux-de Goer, S. et al., The signal transducer PII and bicarbonate acquisition in Prochlorococcus marinus PCC 9511, a marine cyanobacterium naturally deficient in nitrate and nitrite assimilation. *Microbiology* 2002, *148*, 2405–2412.

[28] Herrero, A., Muro-Pastor, A. M., Flores, E., Nitrogen control in cyanobacteria. *J. Bacteriol.* 2001, *183*, 411–425.

[29] Zinser, E. R., Johnson, Z. I., Coe, A., Karaca, E. et al., Influence of light and temperature on Prochlorococcus ecotype distributions in the Atlantic Ocean. *Limnol. Oceanogr.* 2007, *52*, 2205–2220.

[30] Santoro, A. E., Dupont, C. L., Richter, R. A., Craig, M. T. et al., Genomic and proteomic characterization of '*Candidatus Nitrosopelagicus brevis*': an ammonia-oxidizing archaeon from the open ocean. *Proc. Natl. Acad. Sci.* 2015, *112*, 1173–1178.

[31] Dupont, C. L., Rusch, D. B., Yooseph, S., Lombardo, M.-J. et al., Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J.* 2012, *6*, 1186–1199.

[32] Grote, J., Thrash, J. C., Huggett, M. J., Landry, Z. C. et al., Streamlining and core genome conservation among highly divergent members of the SAR11 clade. *MBio* 2012, *3*, e00252-00212.

[33] Yooseph, S., Sutton, G., Rusch, D. B., Halpern, A. L. et al., The Sorcerer II Global Ocean Sampling Expedition: expanding the universe of protein families. *PLOS Biol.* 2007, *5*, 398–431.

[34] Eng, J., McCormack, A., Yates, J., An approach to correlate tandem mass spectral data of peptides with amino acid sequences in a protein database. *J. Am. Soc. Mass Spectrom.* 1994, *5*, 976–989.

[35] Craig, R., Beavis, R. C., A method for reducing the time required to match protein sequences with tandem mass spectra. *Rapid Commun. Mass Spectrom.* 2003, *17*, 2310–2316.

[36] Koonin, E. V., Wolf, Y. I., Karev, G. P., The structure of the protein universe and genome evolution. *Nature* 2002, *420*, 218–223.

[37] Osada, H., Hertweck, C., Exploring the chemical space of microbial natural products. *Curr. Opin. Chem. Biol.* 2009, *13*, 133–134.

