ISME

www.nature.com/ismej

ARTICLE OPEN

Proteomic traits vary across taxa in a coastal Antarctic phytoplankton bloom

J. Scott P. McCain[1,2], Andrew E. Allen[3,4] and Erin M. Bertrand[1,2]✉

Production and use of proteins is under strong selection in microbes, but it is unclear how proteome-level traits relate to ecological strategies. We identified and quantified proteomic traits of eukaryotic microbes and bacteria through an Antarctic phytoplankton bloom using in situ metaproteomics. Different taxa, rather than different environmental conditions, formed distinct clusters based on their ribosomal and photosynthetic proteomic proportions, and we propose that these characteristics relate to ecological differences. We defined and used a proteomic proxy for regulatory cost, which showed that SAR11 had the lowest regulatory cost of any taxa we observed at our summertime Southern Ocean study site. Haptophytes had lower regulatory cost than diatoms, which may underpin haptophyte-to-diatom bloom progression in the Ross Sea. We were able to make these proteomic trait inferences by assessing various sources of bias in metaproteomics, providing practical recommendations for researchers in the field. We have quantified several proteomic traits (ribosomal and photosynthetic proteomic proportions, regulatory cost) in eukaryotic and bacterial taxa, which can then be incorporated into trait-based models of microbial communities that reflect resource allocation strategies.

The ISME Journal (2022) 16:569–579; https://doi.org/10.1038/s41396-021-01084-9

INTRODUCTION

Microbes are constantly faced with an optimization problem: which proteins should be produced, when, and how many? The solutions to this problem dictate metabolic rates, cell stoichiometry, and taxonomic distribution [1–5]. Yet, it is unclear what these solutions actually are in terms of proteome composition, and if different microbes have arrived at different solutions. Microbes are typically compared based on their unique repertoires of potential proteins (e.g., [6–8]), but taxa have shared proteins as well—are these shared proteins produced in similar amounts? Or, do taxa produce distinct amounts under identical conditions? Diverse taxa produce proteins in strikingly similar ratios within some pathways [9], but is stoichiometry conserved between pathways? The answers to these questions will direct future efforts for modeling microbial communities. Perhaps microbes can be represented as collections of genes [10, 11], or, perhaps variation in proteome composition will shed light on the underpinnings of their ecological strategies and biogeochemical contributions.

Ecological strategies are ultimately tied to cellular functions and thus gene expression [12], and models can experimentally test hypotheses to evaluate such connections. Material models (i.e., cultures) have clearly demonstrated that selection acts strongly on protein production [13–15]. While powerful, these approaches are limited to only a few culturable organisms, which can overlook core differences found in less-studied organisms (e.g., [16]). Computational models have also characterized trade-offs and metabolic behaviors in microbes (e.g., [17–19]). While models are critical from a reductionist perspective, characterization and prediction of microbial activity in their environments remains a central research goal.

Observing and measuring gene expression in microbes in situ can also link resource allocation to ecological strategies (e.g., [5, 20–24]). For example, diatom and haptophyte transcriptional dynamics reflect their distinct growth strategies, inferred using metatranscriptomics [22, 23]. Metaproteomics has similarly identified increased abundance of transporter proteins across an oceanographic gradient of decreasing nutrients [5]. Both of these meta-omic approaches can quantify in situ resource allocation, but proteins cost more to produce and therefore better reflect resource allocation [25]. To our knowledge, metaproteomics has not been used to quantify variation in resource allocation strategies across microbial groups.

Our objective was to identify and quantify proteomic "traits" for various eukaryotes and bacteria, by examining microbial proteome composition through a four-week time series at the Antarctic sea ice edge. We define a proteomic trait as a characteristic of an organism at the proteome-level, that includes both the abundance and identity of a protein (or group of proteins), and is connected to organismal fitness or performance [26]. Metaproteomics is confronted by several methodological issues and biases, which we rigorously assess in order to characterize these proteomes. We subsequently provide practical recommendations for researchers using metaproteomics to examine microbial resource allocation. Our analyses suggest

[1] Department of Biology, Dalhousie University, Halifax, NS, Canada. [2] Centre for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University, Halifax, NS, Canada. [3] Microbial and Environmental Genomics, J. Craig Venter Institute, La Jolla, CA, USA. [4] Integrative Oceanography Division, Scripps Institution of Oceanography, University of California San Diego, La Jolla, CA, USA. ✉email: erin.bertrand@dal.ca

Received: 30 April 2021 Revised: 13 July 2021 Accepted: 2 August 2021
Published online: 4 September 2021

examining "coarse-grained" proteomes provides a host of conceptual and technical advantages (coarse-grained defined as a grouping of functionally or taxonomically (Phylum, Class, Order) related proteins). Next we use this approach to connect proteomic resource allocation to the ecology of these plankton. Lastly, we suggest that characterizing coarse-grained proteomes may be useful for assessing nutrient deficiency in the ocean.

METHODS

Field sampling

We collected samples once per week over four weeks at the Antarctic sea ice edge, in McMurdo Sound, Antarctica (December 28, 2014 "GOS-927"; January 6 "GOS-930", 15 "GOS-933", and 22 "GOS-935", 2015; as previously described in [27]). Sea water (150–250 l) was pumped sequentially through three filters of decreasing size (3.0, 0.8, and 0.1 µm, 293 mm Supor filters). Separate filter sets were acquired for metagenomic, metatranscriptomic, and metaproteomic analyses, over the course of ~3 h, each week (36 filters in total). Filters for nucleic acid analyses were preserved with a sucrose-based buffer (20 mM EDTA, 400 mM NaCl, 0.75 M sucrose, 50 mM Tris-HCl, pH 8.0) with RNAlater (Life Technologies, Inc.). Filters for protein analysis were preserved in the same sucrose-based buffer but without RNAlater. Filters were flash frozen in liquid nitrogen in the field and subsequently stored at −80 °C until processed in the laboratory.

Metagenomic and metatranscriptomic sequencing

We used metagenomics and metatranscriptomics to obtain reference databases of potential proteins for metaproteomics. We additionally used a database assembled from a similarly processed metatranscriptomic incubation experiment [28], conducted with source water from the January 15, 2015 time point (these samples were collected on a 0.2 µm Sterivex filter and processed as previously described).

For samples from the GOS-927, GOS-930, GOS-933, and GOS-935 filters, RNA was purified from a DNA and RNA mixture [29]. In total, 2 µg of the DNA and RNA mixture was treated with 1 µl of DNase (2 U/µl; Turbo DNase, TURBO DNase, Thermo Fisher Scientific), followed by processing with an RNA Clean and Concentrator kit (Zymo Research). An Agilent TapeStation 2200 was used to observe and verify the quality of RNA. In total, 200 ng of total RNA was used as input for rRNA removal using Ribo-Zero (Illumina) with a mixture of plant, bacterial, and human/mouse/rat Removal Solution in a ratio of 2:1:1. An Agilent TapeStation 2200 was used to subsequently observe and verify the quality of rRNA removal from total RNA. rRNA-deplete total RNA was used for cDNA synthesis with the Ovation RNA-Seq System V2 (TECAN, Redwood City, USA). DNA was extracted for metagenomics from the field samples (GOS-927, GOS-930, GOS-933, and GOS-935) according to [29]. RNase digestion was performed with 10 µl of RNase A (20 mg/ml) and 6.8 µl of RNase T1 (1000 U/µl), which were added to 2 µg of genomic DNA and RNA mixture in a total volume of 100 µl, followed by 1 h incubation at 37 °C and subsequent ethanol precipitation in −20 °C overnight.

Samples of double stranded cDNA and DNA were fragmented using a Covaries E210 system with the target size of 400 bp. In total, 100 ng of fragmented cDNA or DNA was used as input into the Ovation Ultralow System V2 (TECAN, Redwood City, USA), following the manufacturer's protocol. Ampure XP beads (Beckman Coulter) were used for final library purification. Library quality was analyzed on a 2200 TapeStation System with Agilent High Sensitivity DNA 1000 ScreenTape System (Agilent Technologies, Santa Clara, CA, USA). Twelve DNA and 18 cDNA libraries were combined into two pools with concentration 4.93 and 4.85 ng/µl, respectively. Resulting library pools were subjected to one lane of 150 bp paired-end HiSeq 4000 sequencing (Illumina). Prior to sequencing, each library was spiked with 1% PhiX (Illumina) control library. Each lane of sequencing resulted in between 106,000 and 111,000 Mbp total and 6900–12,000 Mbp and 4800–6900 Mbp for individual DNA or cDNA libraries, respectively.

Metagenomic and metatranscriptomic bioinformatics

Metagenomic and metatranscriptomic data were annotated with the same pipelines. Briefly, adapter and primer sequences were filtered out from the paired reads, and then reads were quality trimmed to Phred33. rRNA reads were identified and removed with riboPicker [30]. We then assembled reads into transcript contigs using CLC Assembly Cell, and then we used FragGeneScan to predict open reading frames (ORFs) [31]. ORFs were functionally annotated using Hidden Markov models and blastp against PhyloDB [32]. Annotations which had low mapping coverage were filtered out (less than 50 reads total over all samples), as were proteins with no blastp hits and no known domains. For each ORF, we assigned a taxonomic affiliation based on Lineage Probability Index taxonomy [32, 33]. Taxa were assigned using two different reference databases: NCBI nt and PhyloDB [32]. Unless otherwise specified, we used taxonomic assignments from PhyloDB, because of the good representation of diverse marine microbial taxa.

ORFs were clustered by sequence similarity using Markov clustering (MCL) [34]. Sequences were assigned MCL clusters by first running blastp for all sequences against each other, where the query was the same as the database. The MCL algorithm was subsequently used with the input as the matrix of *E*-values from the blastp output, with default parameters for the MCL clustering. MCL clusters were then assigned consensus annotations based on KEGG, KO, KOG, KOG class, Pfam, TIGRFAM, EC, GO, annotation enrichment [28, 32, 35–39]. Proteins were assigned to coarse-grained protein pools (ribosomal and photosynthetic proteins) based on these annotations. For assignment, we used a greedy approach, such that a protein was assigned a coarse-grained pool if at least one of these annotation descriptions matched our search strings (we also manually examined the coarse grains to ensure there were no peptides that mapped to multiple coarse-grained pools). For photosynthetic proteins, we included light harvesting proteins, chlorophyll a-b binding proteins, photosystems, plastocyanin, and flavodoxin. For ribosomal proteins, we just included the term "ribosom*" (where the * represents a wildcard character), and excluded proteins responsible for ribosomal synthesis.

Sample preparation and LC-MS/MS

We extracted proteins from the samples by first performing a buffer exchange from the sucrose-buffer to an SDS-based extraction buffer, after which proteins were extracted from each filter individually (as previously described) [27]. After extraction and acetone-based precipitation, we prepared samples for liquid chromatography tandem mass spectrometry (LC-MS/MS). Precipitated protein was first resuspended in urea (100 µl, 8 M), after which we measured the protein concentration in each sample (Pierce BCA Protein Assay Kit). We then reduced, alkylated, and enzymatically digested the proteins: first with 10 µl of 0.5 M dithiothreitol for reduction (incubated at 60 °C for 30 min), then with 20 µl of 0.7 M iodoacetamide (in the dark for 30 min), diluted with ammonium bicarbonate (50 mM), and finally digested with trypsin (1:50 trypsin:sample protein). Samples were then acidified and desalted using C-18 columns (described in detail in ref. [40]).

To characterize each metaproteomic sample, we employed one-dimensional liquid chromatography coupled to the mass spectrometer (VelosPRO Orbitrap, Thermo Fisher Scientific, San Jose, California, USA; detailed in [40]). For each injection, protein concentrations were equivalent across sample weeks, but different across filter sizes. We had higher amounts of protein on the largest filter size (3.0 µm) and less on the smaller filters, so we performed three replicate injections per 3.0 µm filter sample, and two replicate filter injections for 0.8 and 0.1 µm filters. We used a non-linear LC gradient totaling 125 min. For separation, peptides eluted through a 75 µm by 30 cm column (New Objective, Woburn, MA), which was self-packed with 4 µm, 90 A, Proteo C18 material (Phenomenex, Torrance, CA), and the LC separation was conducted with a Dionex Ultimate 3000 UHPLC (Thermo Scientific, San Jose, CA).

LC-MS/MS bioinformatics—database searching, configuration, and quantification

Metaproteomics requires a database of potential protein sequences to match observed mass spectra with known peptides. Because we had sample-specific metagenome and metatranscriptome sequencing for each metaproteomic sample, we assessed various database configurations, including those that we predict would be suboptimal, to examine potential options for future metaproteomics researchers. We used five different configurations, described below. In each case, we appended a database of common contaminants (Global Proteome Machine Organization common Repository of Adventitious Proteins). We evaluated the performance of different database configurations based on the number of peptides identified (using a peptide false discovery rate of 1%).

In order to make these databases (Table 1), we performed three separate assemblies on (1) the metagenomic reads (from samples GOS-927, GOS-930, GOS-933, and GOS-935), (2) metatranscriptomic reads (from samples GOS-927, GOS-930, GOS-933, and GOS-935), and (3) metatranscriptomic

Table 1. Characteristics of the five different database configurations we used for metaproteomic database searches.

Database configuration	Filter size	Number of protein sequences in the database
One-Sample Database	0.1	6.65E + 05
One-Sample Database	0.8	6.42E + 05
One-Sample Database	3	3.34E + 05
Sample-Specific Databases[a]	0.1	7.13E + 05
Sample-Specific Databases[a]	0.8	6.34E + 05
Sample-Specific Databases[a]	3	4.41E + 05
Pooled-across-sizes Databases	0.1	8.37E + 05
Pooled-across-sizes Databases	0.8	8.55E + 05
Pooled-across-sizes Databases	3	7.23E + 05
Metatranscriptome Experiment ($T = 0$)	0.2	4.44E + 05
Metatranscriptome Experiment (all)	0.2	2.19E + 06

For the "One-Sample Database", the first time point was used, and all samples were matched according to filter sizes. For the "Sample-Specific Databases", each database was matched with the corresponding metaproteomic sample. For the "Pooled-Across-Sizes Databases", databases were pooled across every time point and matched according to filter size. For these aforementioned databases, the metagenomic and metatranscriptomic protein coding sequences were pooled. For the "Metatranscriptome Experiment ($T = 0$)", only the first sampling point from the metatranscriptome experiment was included. For the "Metatranscriptome Experiment (all)" configuration, all protein coding sequences were included from the treatment outcomes as well as the $T = 0$.

[a]Averages are presented for Sample Specific Databases.

reads from a concurrent metatranscriptomic experiment, started at the location where GOS-933 was taken [28]. Database configurations were created by subsetting from these assemblies. The first configuration was "one-sample database", constructed to represent the scenario where only one sample was used for metagenomic and metatranscriptomic sequencing (we chose the first sampling week). Specifically, this was done by subsetting and including ORFs from the metagenomic and metatranscriptomic assemblies if reads from this time point were present in that sample (reads mapped as in [28]), and then removing redundant protein sequences (P. Wilmarth, fasta utilities). The second configuration was the "sample-specific database", where each metaproteomic sample had one corresponding database (prepared from both metagenome and metatranscriptome sequencing completed at the same sampling site), also done by subsetting ORFs from the metagenomic and metatranscriptomic assemblies as described above. The third configuration was pooling databases across size fractions—such that all metagenomic and metatranscriptomic sequences across the same filter sizes (e.g., 3.0 µm) were combined. ORFs were subsetted from the metagenomic and metatranscriptomic assemblies as above. The fourth and fifth configurations are from the concurrent metatranscriptomic experiment [28]. The fourth configuration ("metatranscriptome experiment (T0)") was the metatranscriptome of the in situ microbial community (i.e., at the beginning of the experiment). This database was created by subsetting from the "metatranscriptome experiment (all)" assembly. Finally, the fifth configuration was the metatranscriptome of all experimental treatments pooled together (two iron levels, three temperatures; "metatranscriptome experiment (all)"). The overlap between databases (potential tryptic peptides) in different samples is presented graphically in Supplementary Figs. S1–S3.

After matching mass spectra with peptide sequences for each database configuration (MSGF + with OpenMS, with a 1% false discovery rate at the peptide level; [41, 42]), we used MS1 ion intensities to quantify peptides. Specifically, we used the FeatureFinderIdentification approach, which cross-maps identified peptides from one mass spectrometry experiment to unidentified features in another experiment—increasing the number of peptide quantifications [43]. This approach requires a set of experiments to be grouped together (i.e., which samples should use this cross-mapping?). We grouped samples based on their filter sizes (including those samples that are replicate injections). First, mass spectrometry runs within each group were aligned using MapAlignerIdentification [44], and then FeatureFinderIdentification was used for obtaining peptide quantities.

After peptides have been identified and quantified, we mapped them to proteins or MCL clusters of proteins, which have corresponding functional annotations (KEGG, KO, KOG, Pfams, TIGRFAM; [28, 32, 35–39]). Functional annotations were used in three separate analyses. (1) Exploring the overall functional changes in microbial community metabolism, we mapped peptides to MCL clusters—groups of proteins with similar sequences. These clusters have consensus annotations based on the annotations of proteins found within the clusters (described in detail in [28]). For this section, we only used peptides that uniquely map to MCL clusters. (2) We restricted the second analysis to two protein groups: ribosomal and photosynthetic proteins. For this analysis, we mapped peptides to one of these protein groups if at least one annotation mapped to the protein group (via string matching with keywords). This approach is "greedy" because does not exclude peptides if they also correspond with other functional groupings, but this is necessary because of the difficulties in comparing various annotation formats. (3) The last analysis for functional annotations was for targeted proteins, and we only mapped functions to peptides where the peptides uniquely identify a specific protein (e.g., plastocyanin).

Code for the database setup and configuration, database searching, and peptide quantification is open source (https://github.com/bertrand-lab/ross-sea-meta-omics).

LC-MS/MS bioinformatics—normalization

Normalization is an important aspect of metaproteomics: it influences all inferred peptide abundances. Typically, the abundance of a peptide is normalized by the sum of all identified peptide abundances. We use the term normalization factor for the inferred sum of peptide abundances. Note that the apparent abundance of observed peptides is dependent on the database chosen. In theory, if fewer peptides are observed because of a poorly matching database, this will decrease the normalization factor, and those peptides that are observed will appear to increase in abundance. It is not known how much this influences peptide quantification in metaproteomics.

For each database configuration, we separately calculated normalization factors. We then correlated the sum of observed peptide abundances with each other. To get a database-independent normalization factor, we used the sum of total ion current (TIC) for each mass spectrometry experiment (using pyopenms; [45]), and also examined the correlation with database-dependent normalization factors. If normalization factors are highly correlated with each other, that would indicate database choice does not impact peptide quantification. Using TIC for normalization may have drawbacks, particularly if there are differences in contamination, or amounts of non-peptide ions across samples.

Defining proteomic mass fraction

Protein abundance can be calculated in two ways: (1) the number of copies of a protein (independent of a proteins' mass), or (2) the total mass of the protein copies (the sum of peptides). We refer to the latter as a proteomic mass fraction. For example, to calculate a diatom-specific, ribosomal mass fraction, we sum all peptide abundances that are diatom- and ribosome-specific, and divide by the sum of peptide abundances that are diatom-specific. Note that this is slightly different to other methods, like the normalized spectral abundance factor, which normalizes for total protein mass (via protein length; [46]).

Combining estimates across filter sizes

Organisms should separate according to their sizes when using sequential filtration with decreasing filter pore sizes. In practise, however, organisms can break because of pressure during filtration, and protein is typically present for large phytoplankton on the smallest filter size and vice versa. We used a simple method for combining observations across filter sizes, weighted by the number of observations per filter. We begin with the abundance of a given peptide, which was only considered present if it was observed across all injections of the same sample. We calculated the sum of observed peptide intensities (i.e., the normalization factor), and divided all peptide abundances by this normalization factor. Normalized peptide abundances are then averaged across replicate injections. If we are estimating the ribosomal mass fraction of the diatom proteome, we first normalize the diatom-specific peptide intensities as a proportion of diatom biomass (i.e., divide all diatom-specific peptides by the sum of all diatom-specific peptides). We then summed all diatom-normalized peptides intensities that are unique to both diatoms and ribosomal proteins, which would give us the ribosomal proportion of the diatom proteome. Yet, we typically would obtain multiple estimates of, for example, ribosomal mass fraction of diatoms, on different filters. We combined the three values by multiplying each by a coefficient that represents a weight for each observation (specific to a filter size). These coefficients sum to one, and are calculated by summing the total number of peptides observed at a time point for a filter, and dividing by the total number of peptides observed across filters (but within each time point). For example, if we observed 100 peptides that are diatom- and ribosome-specific, and 90 of these peptides were on the 3.0 µm filter and only ten were on the 0.8 µm filter, we would multiply the 3.0 µm filter estimate by 0.9 and the 0.8 µm filter by 0.1. This method uses all available information about proteome composition across different filter sizes (similar to [47]).

When we estimate the proteomic mass fraction of a given protein pool, we do not need to adjust for the total protein on each filter. This is because this measurement is independent of total protein. However, for merging estimates of total relative abundance of different organisms across filters, we needed to additionally weight the abundance estimate by the amount of protein on each filter. Therefore, in addition to the weighting scheme described above, we multiplied taxon abundance estimates by the total protein on each filter divided by the total protein across filters on a given day.

LC-MS/MS simulation

We used simulations of metaproteomes and LC-MS/MS to (1) quantify biases associated with inferring coarse-grained proteomes from metaproteomes, and (2) to mitigate these biases in our inferences. Specifically, we asked the question: how does sequence diversity impact quantification of coarse-grained proteomes from metaproteomes? Consider a three organism microbial community. If two organisms are extremely similar, there will be very few peptides that can uniquely map to those organisms, resulting in underestimated abundance. The third organism would also be underestimated, but to a lesser degree, unless it had a completely unique set of peptides. A similar outcome is anticipated with differences in sequence diversity across protein groups, such that highly conserved protein groups will be underestimated.

Our mass spectrometry simulations offer a unique perspective on this issue: we know the “true” metaproteome, and we can compare this with an “inferred” metaproteome. We simulated variable numbers of taxonomic groups, each with different protein pools of variable sequence diversity. From this simulated metaproteome, we then simulated LC-MS/MS-like sampling of peptides. Complete details of the mass spectrometry simulation are available in [48] and the Supplementary materials. The only difference between this model and that presented in [48] is here we include dynamic exclusion. The ultimate outcomes from these simulations were (1) identifying which circumstances lead to biased inferences about proteomic composition, and (2) determining the underpinnings of these biases.

Cofragmentation bias scores for peptides

We recently developed a computational model (“cobia”) that predicts a peptides’ risk for interference by sample complexity (more specifically, by cofragmentation of multiple peptides; [48]). This study showed that coarse-grained taxonomic and functional groupings are more robust to bias, and that this model can also be used to estimate bias. We ran cobia with the sample-specific databases, which produces a “cofragmentation score”—a measure of risk of being subject to cofragmentation bias. Specifically, the retention time prediction method used was RTPredict [49] with an “OLIGO” kernel for the support vector machine. The parameters for the model were: 0.008333 (maximum injection time); 3 (precursor selection window); 1.44 (ion peak width); and 5 (degree of sparse sampling). Code for running this analysis, as well as the corresponding input parameter file, is found at https://github.com/bertrand-lab/ross-sea-meta-omics.

Description of previously published datasets analyzed

We leveraged several previously published datasets to compare our metaproteomic results. Specifically, we used proteomic data of phytoplankton cultures of *Phaeocystis antarctica* and *Thalassiosira pseudonana* [27, 50], and of cultures of *Escherichia coli* under 22 different culture conditions [51]. Coarse-grained proteomic estimates were also compared with previously published targeted metaproteomic data [27].

RESULTS AND DISCUSSION

We characterized proteomic traits of eukaryotic and bacterial taxa at the Antarctic sea ice edge. To do so, we have leveraged a combination of sample-specific nucleic acid sequencing and metaproteomics, assessing various assumptions and challenges with metaproteomics. Below, we first discuss our methodological results, and then we examine observations of different proteomic traits across microbial taxonomic groups. Finally, we touch on using coarse-grained protein pools for measuring nutrient stress in the ocean.

Database choice influences peptide identifications and quantification

The sequence databases from the metatranscriptome experiment conducted on our third sampling week (January 15, 2015) outperformed sample-specific databases and other configurations (in terms of number of peptide spectrum matches, Supplementary Fig. S4 and Table 1). Specifically, we identified 14,455 unique peptides using the “metatranscriptome experiment T0” database, while 8022 unique peptides were identified with the “sample-specific database” (Supplementary Fig. S4). We identified a core set of 5127 peptides, regardless of the database chosen (Supplementary Fig. S4). The database pooled across time points identified more peptides than the “sample-specific database”, similar to previous work [52]. The metatranscriptomic experiment (both “metatranscriptomic experiment (T0)” and “metatranscriptomic experiment (all)”) were more valuable in identifying larger, primarily eukaryotic organisms (Supplementary Figs. S4–S7). Overall, the two metatranscriptomic experiment databases performed similarly in terms of number of identified peptides. All subsequent analyses use the identified peptides from the “metatranscriptome experiment (all)” database. Importantly, a difference between the metatranscriptomes of sample-specific filters and the metatranscriptomic experiment databases was sequencing depth (Supplementary Table 1). This difference likely influenced the metatranscriptomic read assembly, improving the assembly of eukaryotic-protein sequences and therefore creating a better database (i.e., in terms of peptides identified). Note that databases were constructed after assembly, and then subsetted to create individual databases (Methods). Overall, deep metatranscriptomic sequencing appears to be a promising avenue for metaproteomics with tailored databases (Supplementary Table 1).

Database choice influenced peptide quantification due to normalization. We quantified this by correlating sample-specific normalization factors with each other and with the TIC (i.e., a database-independent normalization, Supplementary Figs. S8–S10). Examining the correlation between the best- and worst-performing databases, there was a range of R^2 values, from 83 to 99% (Supplementary Figs. S8–S10). If we consider a peptide observed in a mass spectrometry experiment with an intensity value of 100, we expect variation in the inferred value to range from 92 to 108, reflecting variation in the normalization factor of 16%. This has significant consequences for comparative metaproteomics: consider two samples, one with a perfectly matched database and a second that uses the same database, but is poorly matched. Using standard methods, the peptides identified in the second sample will appear to increase in abundance, even if the abundance is constant. We anticipate that database choice would similarly affect quantification when other mass spectrometry methods are employed (e.g., labeled untargeted metaproteomics, or experiments using data-independent acquisition). Note that our worst-performing database was still well-matched to the community, so for researchers studying very distinct communities it is vital to address this issue by using database-independent normalization, or ensuring bias across samples is minimal. We provided methods for doing both.

One simple alternative is to use a database-independent metric of total peptide abundance: total MS1 ion current (TIC). We found that TIC is well correlated with the total peptide abundance inferred from the best-performing database (with correlation coefficients 0.98, 0.95, and 0.91 for the 3.0, 0.8, and 0.1 µm filter sizes, Supplementary Figs. S8–S10, respectively). This result has two consequences: (1) it suggested that TIC may be a viable alternative for normalization in comparative metaproteomics. (2) It validated the use of our best-performing database, as we identified most of the abundant peptides in our sample. Given that these two approaches were highly correlated, we used the "metatranscriptome experiment (all)" database for all subsequent analyses.

Taxonomic and functional composition shifted through the season at the Antarctic sea ice edge

Taxonomic abundance shifted through the season at the Antarctic sea ice edge (Fig. 1a). The microbial community was dominated by *Phaeocystis antarctica* (Haptophyta) early in the season, with diatoms increasing in relative abundance later (predominantly *Fragilariopsis* sp. and *Pseudonitzschia* sp.). The phytoplankton bloom progression and high dinoflagellate biomass contribution were both consistent with previous observations in the Ross Sea [53, 54]. Bacterial taxa had relatively lower protein biomass and more consistent relative biomass values through time compared with eukaryotic taxa. Of the bacterial taxa we observed, Rhodobacterales was the most abundant group, with abundances being mostly stable though the season.

We identified shifts in protein abundance by mapping peptides to de novo protein clusters (irrespective of taxonomic assignments)—including protein clusters with no known function. Earlier in the season there was a high relative abundance of Chlorophyll A-B binding proteins and ATP synthase alpha/beta family proteins (Fig. 1b), which is anticipated because of the higher levels of dissolved iron [27]. Demonstrating the importance of de novo protein group assignment, the most abundant protein group in our entire dataset had no functional annotations (Fig. 1b, Unknown Protein Cluster 2818, mostly belonging to Ciliates). Further examination of a representative protein sequence within this cluster found no functionally similar proteins within the NCBI non-redundant database. We suggest that these unknown, highly abundant proteins should be targets for functional characterization.

Eukaryotic and bacterial taxa have taxon-specific proteomic allocation strategies

We quantified two simple proteomic traits of microbes: the ribosomal protein mass fraction and the photosynthetic protein mass fraction (using a combined estimate across filter sizes, Supplementary Fig. S11). Eukaryotic taxa formed unique clusters based on these two traits, with more variation across taxa than across time points (Fig. 2a). For example, haptophytes had relatively high proportions of both ribosomal and photosynthetic protein fractions. Examining the five most abundant bacterial taxa, we also observed distinct proteomic compositions, with Gammaproteobacteria exhibiting the highest ribosomal protein mass fraction (Fig. 2).

Before examining the underpinnings of these proteomic traits, we first scrutinized these inferences using mass spectrometry simulations and additional data sources. Our analysis was limited to coarse-grained protein functions and taxa, which is robust to bias arising from variable sample complexity [48]. Our mass spectrometry simulations suggested that low sequence diversity in taxa or protein groups can lead to underestimation (Supplementary Fig. S12), but this bias is mitigated by examining abundant proteins or taxa. Identifying ~50 peptides or more is evidence that there is sufficient sequence diversity in a protein group to avoid this type of underestimation (Supplementary Fig. S12). We identified greater than 50 taxon-specific peptides for each protein group, indicating that these observations are not subjected to significant biases arising from sequence diversity. We therefore restricted our analyses to taxa and protein groups that are relatively abundant. Note that for dinoflagellates, we observed relatively few peptides in the photosynthetic proteomic mass fraction, so our observations are likely underestimating the true value (Supplementary Discussion, Supplementary Fig. S12). Despite this underestimation, the true value is probably quite low (discussed below).

We provided two additional estimates of ribosomal and photosynthetic protein mass fraction from cultured phytoplankton (Fig. 2c). Metaproteomics can underestimate taxon-specific protein mass when taxonomically uninformative peptides are not used. For example, we might identify a highly conserved peptide produced by a diatom, but are unable to map it to diatoms because it also corresponds to other taxa, and this peptide would be excluded from the quantification of diatoms. Therefore, we compared the ratios of ribosomal to photosynthetic protein mass fraction from the metaproteomic observations to cultured diatoms and haptophytes. Ratios were similar in cultures compared to populations sampled in situ (Fig. 2c), despite such culturing experiments occurring under different environmental conditions. Trends observed in the proportion of transcripts mapped to ribosomal proteins in different groups of bacteria also mirrored our estimates of ribosomal protein mass fraction (high for Gammaproteobacteria and low for SAR11; [21]).

We examined coarse-grained taxonomic groups. It is possible that within these coarse groupings, different taxa included in these groupings employ different allocation strategies. We therefore sought to determine whether taxonomic sub-groupings displayed similar expression patterns. This issue is challenging to assess, because as subgroupings are further examined, there is increased susceptibility to several biases (as outlined above). We therefore examined one subgrouping, diatoms, that contained two dominant species: *Fragilariopsis* sp. and *Pseudonitzschia* sp. The taxonomic assignments for these two diatoms were from the NCBI nt database. We observed similar proteome estimates for both ribosomal and photosynthetic proteins amongst both these subgroups of diatoms (Supplementary Fig. S13), suggesting they are functionally similar based on these proteomic traits. However, we cannot exclude the possibility that for other taxonomic groups the trends observed are due to a diversity of underlying microbial strategies. Yet at this coarse taxonomic level, we concluded that different microbial taxa exhibited distinct coarse-grained proteomes.

We now turn to the ecological relevance of these protein expression patterns. Protein synthesis is the primary energy sink in cells [25], and photosynthesis or respiration is the primary energy source in cells. Why do dinoflagellates have relatively low photosynthetic protein mass fractions? This taxonomic group is typically mixotrophic or heterotrophic [55], which would require larger investment in respiratory proteins for energy production. Haptophytes and diatoms had similar amounts of photosynthetic proteins, but very different amounts of ribosomal proteins (Fig. 2a), so there was no direct trade-off between producing ribosomal versus photosynthetic machinery (i.e., they do not form Pareto front) [56, 57]. Gammaproteobacteria had the highest ribosomal mass fraction within the observed bacterial taxa, and haptophytes had higher ribosomal mass fractions compared to diatoms. Gammaproteobacteria ribosomal mass fraction decreased through the season, perhaps corresponding with a decreased growth rate as micronutrients are depleted by the phytoplankton bloom.

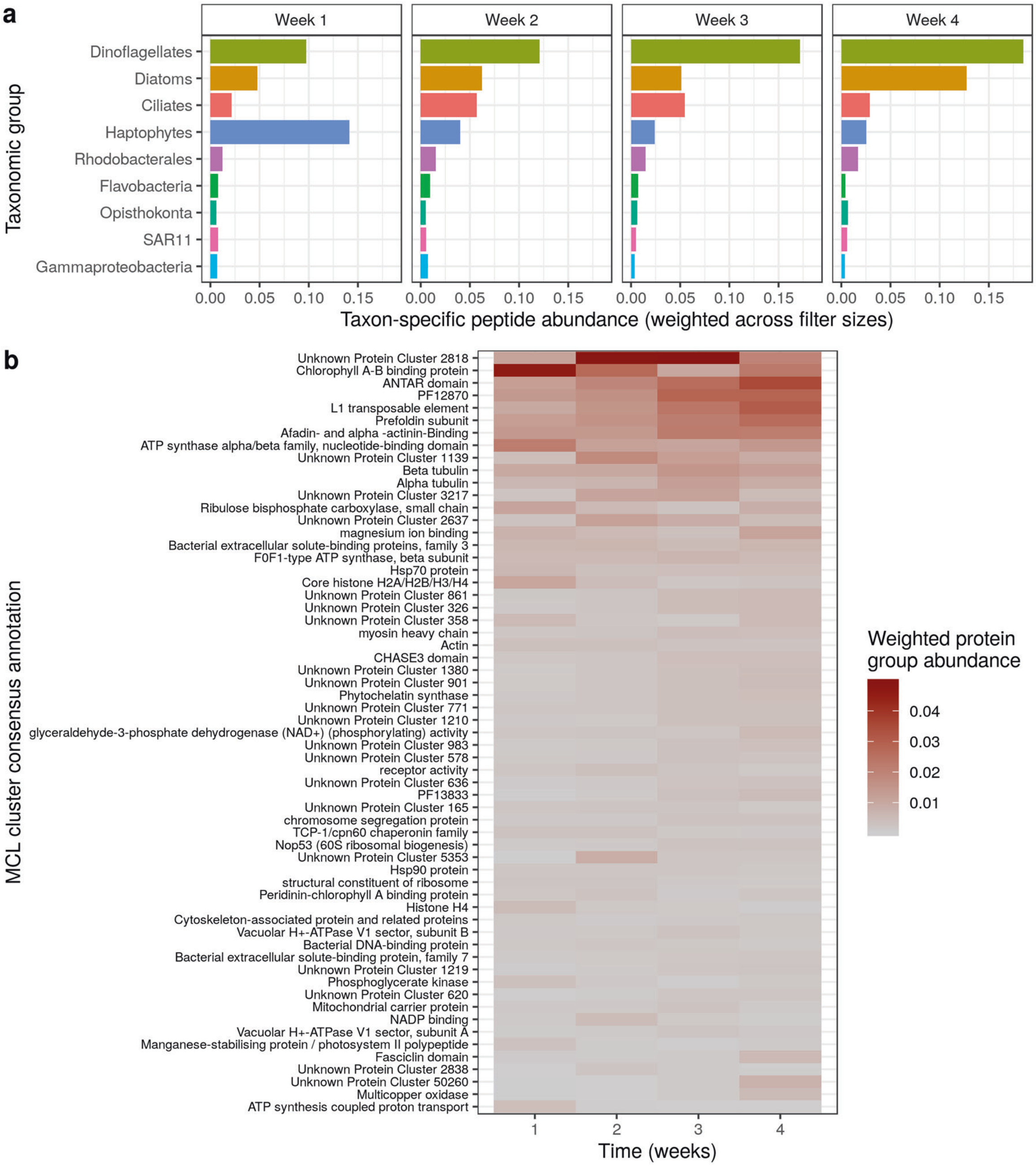

Fig. 1 Taxonomic and functional composition shifted through the season. a Measurements of relative change in protein biomass identified a taxonomic shift at the Antarctic sea ice edge. Protein biomass is calculated as the sum of taxon-specific peptide intensities, weighted by the protein mass per filter for each sampling time. **b** Relative change in protein functional clusters shows that unknown protein clusters contribute greatly to in situ protein biomass, and also identifies a functional shift across weeks.

What are the ecological implications of having more ribosomes? If we assume constant translation rate per translational apparatus (but see [58]), taxa then had different total protein synthesis output. Growth rate is directly related to total protein synthesis output, because protein comprises a large portion of cell mass. To have a faster growth rate, microbes' need to increase protein synthesis (see [12], for derivation and assumptions). We hypothesize that high total protein synthesis output (via high ribosomes) is more advantageous under high nutrient regimes, as it would allow an elevated growth rate. Indeed, haptophytes and Gammaproteobacteria were more abundant earlier in the season (which had higher concentrations of dissolved Fe and Mn) [27]. Another interpretation is that these early-abundant taxa are better suited to a dynamic environment. Perhaps these early-abundant taxa (Gammaproteobacteria, haptophytes) increased investment in ribosomes as a form of bet hedging, which enables a faster growth rate in a dynamic environment [59].

Environment-independent proteomic fraction varies across taxa

What is the cost of responding quickly to a dynamic environment? We hypothesized that there is a regulatory cost for producing proteins that are optimal for a set of environmental conditions. Constitutive protein production does not incur this regulatory cost at the risk of being mismatched to environmental conditions. If the proteome is mostly constant across conditions, this indicates a

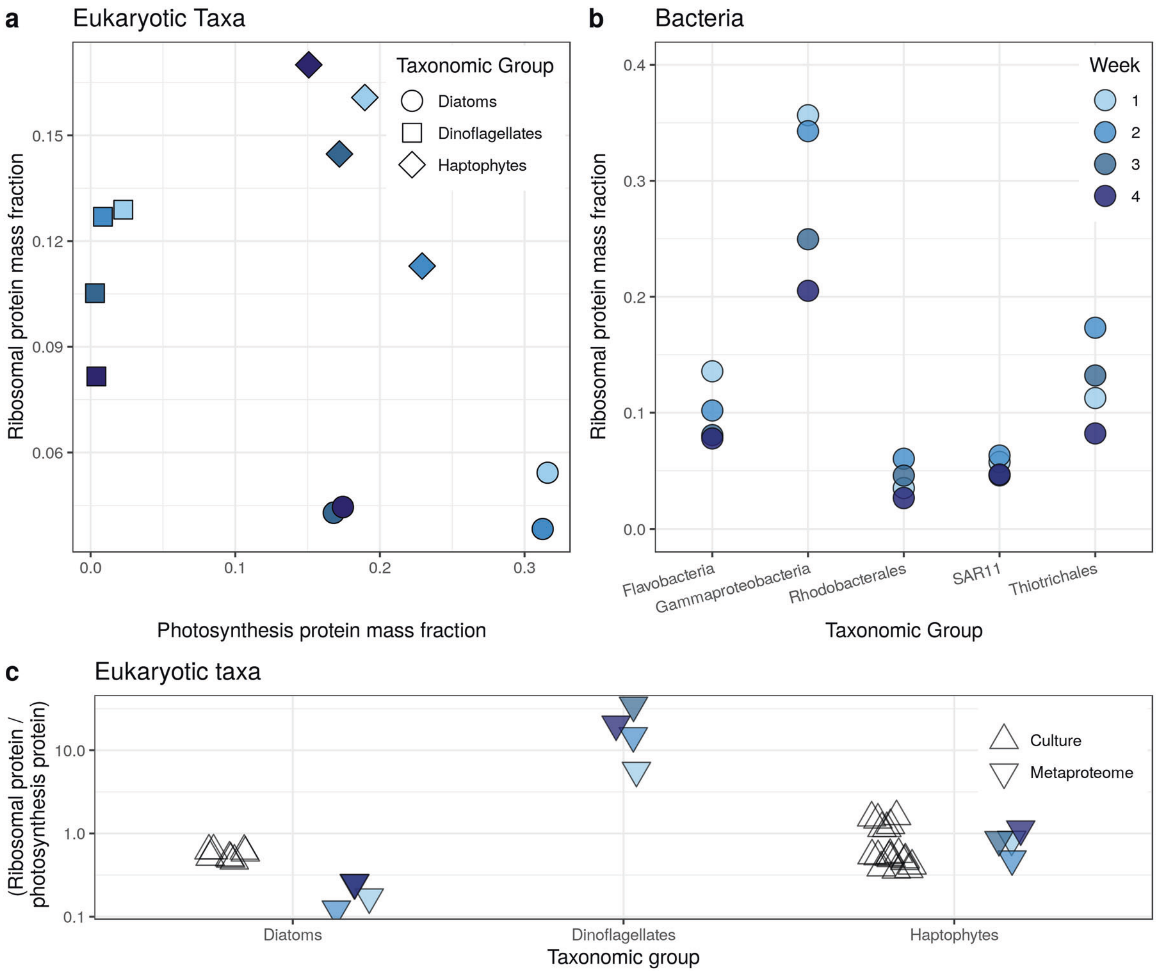

Fig. 2 Eukaryotic and bacterial taxa have taxon-specific proteomic allocation strategies. a Photosynthetic protein mass fraction and ribosomal protein mass fraction (normalized by total amount of a given taxon at a each time point) identifies clear taxonomic subgroupings. **b** Examining bacteria only shows variation in ribosomal mass fraction across groups. Note that Thiotrichales are an order within Gammaproteobacteria, so Gammaproteobacteria here refers to all non-Thiotrichales Gammaproteobacteria. **c** Ratios of ribosomal protein mass fraction to photosynthetic protein mass fraction derived from metaproteomic observations and compared with phytoplankton proteomes observed in culture (*Phaeocystis antarctica*, *Thalassiosira pseudonana* [27, 50]).

low regulatory cost, and vice versa. We propose a proteomic trait that reflects regulatory cost: the proteomic fraction that is environment-independent. This proteomic trait is quantifiable using metaproteomics, and due to the dynamic nature of the ocean, is likely an important selective force for marine microbes.

We classified peptides that are relatively constant across different environmental conditions, and then summed their average intensities to get an environment-independent peptide mass fraction (Fig. 3b, c). Note that (1) peptide intensities were first normalized by total taxon-specific peptide intensity (they therefore sum to one for each taxon), and (2) estimates of environment-independent peptide mass fraction were combined across filter sizes. Using previously published proteomic data from replicate cultures of *E. coli* under identical conditions, we chose a cut-off point distinguishing environment-dependent versus -independent peptides (represented with vertical lines, Fig. 3a and Supplementary Fig. S14) [51]. This cut-off point was calculated by examining the distribution of protein-level coefficients of variation for each *E. coli* culture condition, determining the third quartile, and then taking the mean across all culture conditions [51]. We then can determine the proportion of the proteome that is environment-independent and -dependent (using the mean abundance value per peptide). There are potential biases in this novel method. We address the impact of these biases using published data and by making comparisons with other estimates of regulatory costs across taxa from previously published work (see Supplementary Discussion, Supplementary Figs. S14 and S15).

SAR11 had the highest environment-independent peptide mass fraction across all eukaryotic and bacterial taxa we examined (Fig. 3a, b and Supplementary Fig. S16), consistent with previous work suggesting SAR11 has reduced regulatory investment [60]. Within eukaryotes, dinoflagellates exhibited the highest environment-independent peptide mass fraction, and dinoflagellates in other oceanic regions also exhibited lower regulatory cost [20, 22].

Diatoms had a lower environment-independent proteomic fraction compared with haptophytes, suggesting they have higher regulatory costs. Recall the previous result that diatoms had a lower proportion of ribosomes compared with haptophytes (but similar proportions of photosynthetic proteins; Fig. 2a). We speculate that two proteomic traits comprise a trade-off for these two taxa: higher total protein synthesis via more ribosomes (i.e., leading to fast growth under high nutrient conditions), but at a cost of being less able to dynamically regulate their proteomes. This suggests that in a high nutrient environment (that is also dynamic), dynamically responding to the environment is not the optimal strategy. Instead, a better strategy is constitutively expressing proteins that are favorable for rapid growth (e.g., high

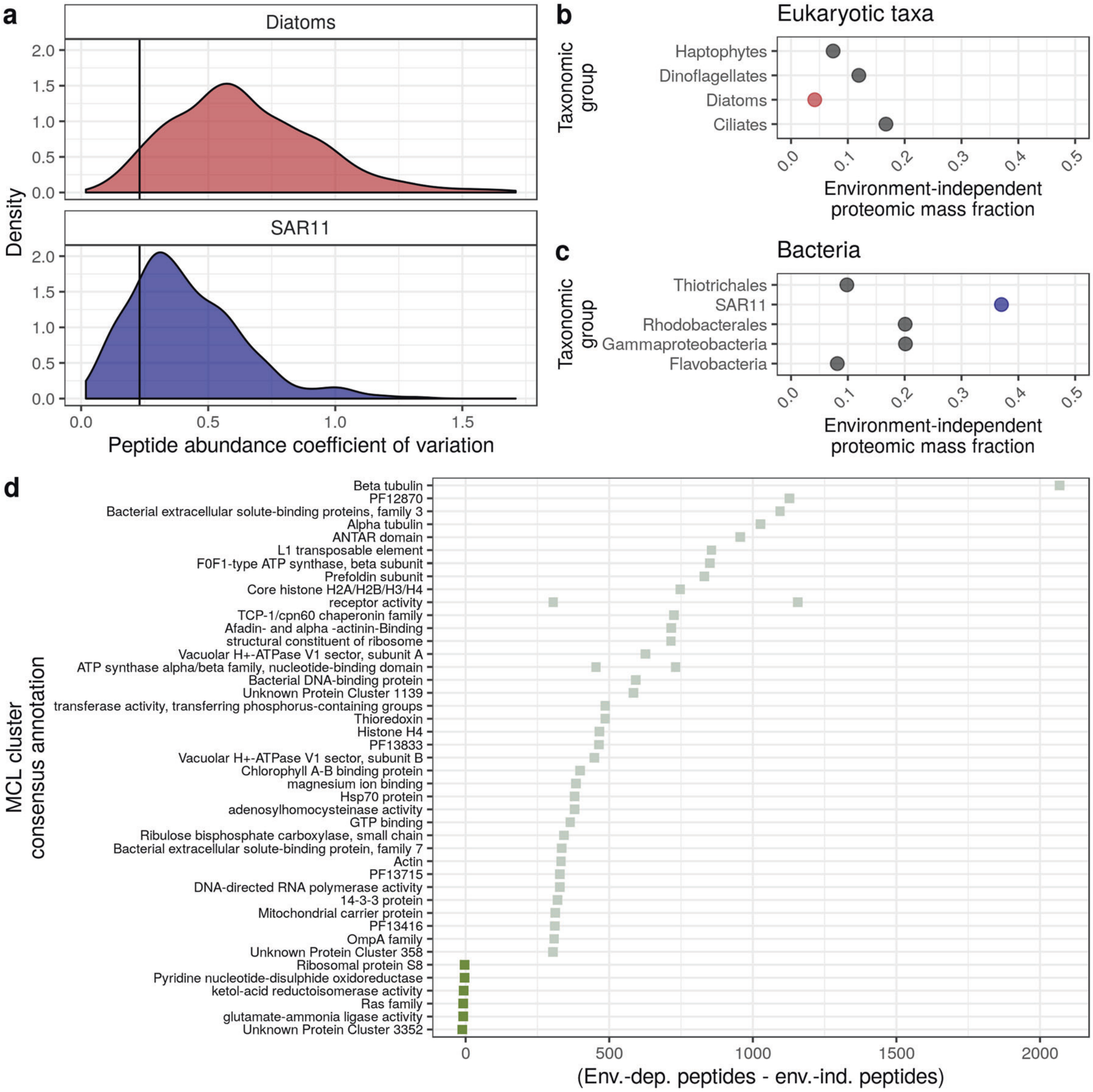

Fig. 3 Environment-independent proteomic fraction varies across taxa and functional protein groups. a The distribution of peptide-specific coefficients of variation can be used to identify if a peptide is significantly changing across environmental conditions. Peptide abundance is first divided by the total taxon-specific peptide intensity. Diatoms and SAR11 represent two extremes within this dataset—diatoms have a highly variable proteome while SAR11 has a relatively constant protein expression. The cutoff point was chosen using replicate cultures of *E. coli* protein expression (vertical line, [51]). **b**, **c** After classifying peptides by their coefficients of variation, we categorized peptides as independent of their environment and those that are not. Points represent the sum of peptide intensities that are environment-independent across eukaryotic and bacterial taxa. **d** A comparison of protein functional clusters that have peptides classified mostly as environment-dependent or environment-independent. The values plotted are the number of environment-dependent peptides minus the number of environment-independent peptides. Note that some MCL clusters had the same consensus annotation (e.g., receptor activity), which is why there are multiple points on some horizontal lines. Positive values indicate that more peptides observed within this protein cluster were dependent on their environment (shown in light green), while negative values indicating more peptides within this protein cluster were identified as environment-independent (shown in dark green).

ribosomal production in haptophytes). The lower ribosomal mass fraction observed in diatoms would limit their growth in higher micronutrient environments but make them more successful in lower micronutrient environments. Ross Sea phytoplankton blooms typically progress from haptophyte- to diatom-dominated, as micronutrients stocks (e.g., Fe and Mn) transition from replete to deplete [53, 61–63]. There is also evidence that *Phaeocystis* has a higher Fe requirement [64], which may be related to these proteomic traits. We posit that differences in regulatory cost and ribosomal mass fraction between diatoms and haptophytes may help explain their ecological succession.

Are some protein functions more often categorized as environment-independent or environment-dependent? Highlighting some examples, the actin protein cluster was often classified as environment-dependent (Fig. 3c). Actin is involved in endocytosis, and inorganic Fe uptake occurs via an endocytotic mechanism (with phytotransferrin) [65]. Perhaps variable expression of actin is related to the amount of bioavailable Fe, and previously published proteomic experiments also showed that actin was differentially expressed due to Fe [66, 67]. ATP synthase-peptides and chlorophyll A-B binding protein-peptides were also mostly classified as environment-dependent, likely reflecting

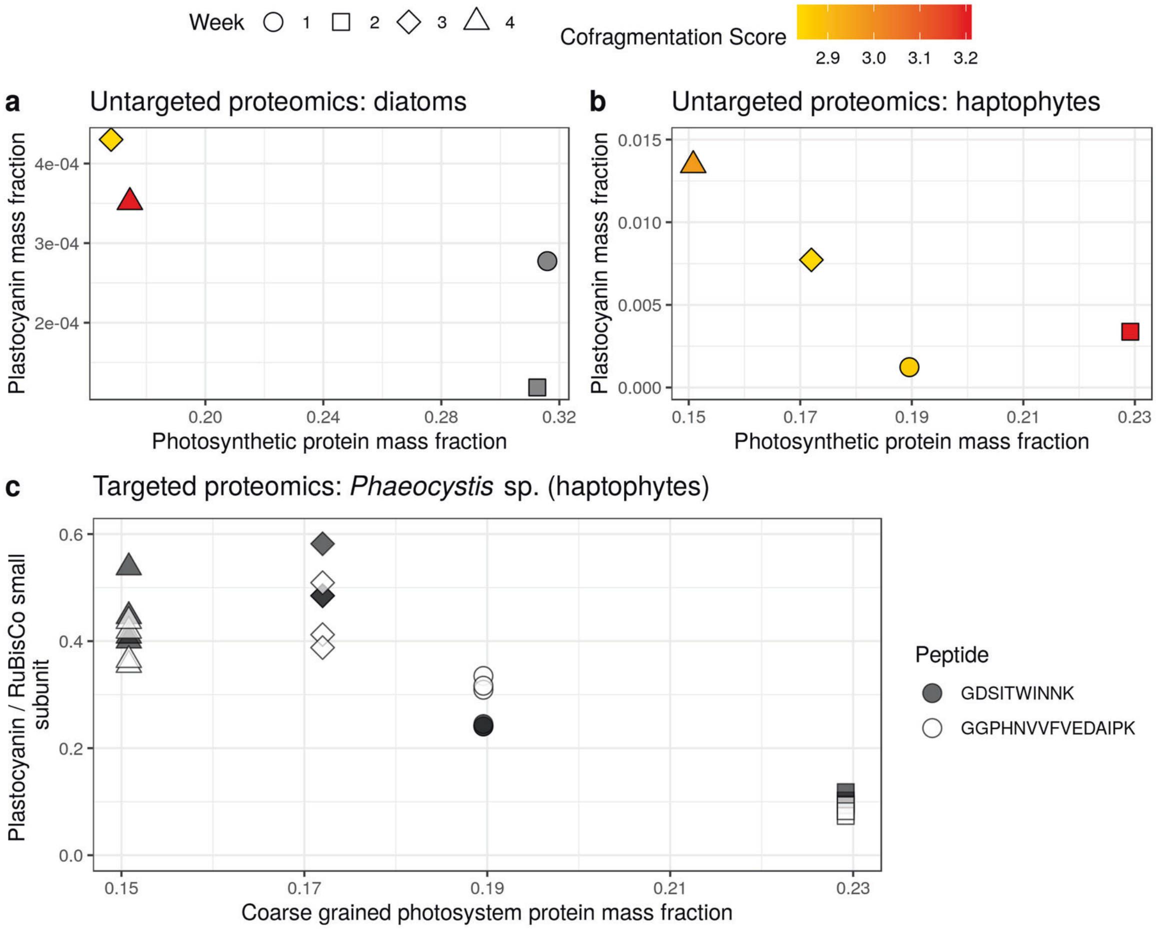

Fig. 4 Coarse-grained proteomes can be used to assess nutrient stress. a, **b** Comparison of the single-protein biomarker plastocyanin with the photosynthetic protein mass fraction for diatoms and haptophytes (using discovery proteomics). Points are colored with their corresponding, sample-specific cofragmentation score (the number of potentially cofragmenting peptides). Cofragmentation scores were calculated using the sample-specific nucleic acid sequencing, and points colored in gray correspond to peptides that were identified and quantified with the “Metatranscriptome experiment (all)” database, but were not present in the sample-specific databases. **c** Comparison of the single-protein biomarker (using targeted proteomics) plastocyanin with the photosynthetic protein mass fraction for haptophytes. Two peptides for plastocyanin are shown, and each point represents one technical replicate measurement. *Phaeocystis* plastocyanin abundance is normalized to Phaeocystis RuBisCO small subunit abundance, where we used the mean of two taxon-specific peptides (AKPNFYVK and QIQYALNK) to calculate RuBisCO abundance [27].

higher primary production earlier in the season (Fig. 3c). In contrast, the ketol-acid reductoisomerase protein cluster (involved in branched-chain amino acid synthesis) was mostly classified as environment-independent. It is unclear what the mechanistic basis for constitutive expression of this protein might be, but several proteomic studies of diatoms also suggest similar expression across conditions [50, 66, 67]. Using this extensible approach to identify constitutively expressed proteins across a wide array of taxa would shed light on these mechanisms. With vastly more metaproteomic data being generated (e.g., [68]), identifying constitutively expressed proteins across diverse taxa would help answer the question: what are the features of constitutively expressed proteins? For example, perhaps there are certain protein functional groupings that are often constitutively expressed.

Coarse-grained proteomes can assess nutrient stress

Proteomics is also used in marine microbiology to assess stress corresponding to a deficient nutrient (e.g., [4]). For example, expression of the protein plastocyanin may reflect Fe deficiency, because plastocyanin does not contain Fe and performs a similar function as the Fe-containing protein cytochrome *c* [69]. Biomarkers of physiological stress are increasingly nuanced (e.g., [27]), sometimes taxon specific, and can require targeted mass spectrometry approaches. Coarse-grained approaches may be a complementary method for assessing stress or nutrient deficiency. We compared using coarse-grained proteomes with single-protein biomarkers. Previous bottle incubation work and targeted metaproteomics showed that there was a transition to Fe- and Mn stress at this sampling location in the Ross Sea, so we focus on Fe-stress indicators [27]. We first solely examined the photosynthetic protein mass fraction compared to the mass fraction of peptides assigned to the plastocyanin, for diatoms and haptophytes (Fig. 4a, b). This approach is biased by variable complexity across samples [48], but we predicted the degree of bias with a quantitative metric (the “cofragmentation score”). This score reflects the expected number of peptides with similar *m/z* and retention times. Overall, there were relatively few potential cofragmenting peptides (≈3), indicating low bias (peptides with high bias can have upwards to 300 cofragmenting peptides, for example [48]). We observed a negative relationship between the photosynthetic protein mass fraction and the plastocyanin mass fraction (note that these two variables are not independent, as plastocyanin is considered as part of the photosynthetic mass fraction). We also examined *Phaeocystis antarctica*-specific peptides measured with previously published targeted mass spectrometry, and identified a negative correlation between the abundance values of plastocyanin and the coarse-grained estimates of photosynthetic proteins (Fig. 4c). We conducted this analysis as a proof-of-concept for using coarse-grained proteomes to assess nutrient deficiency, as coarse-grained proteomes are amenable for untargeted metaproteomic analyses. These preliminary analyses suggest that coarse-grained proteome composition may be a useful tool for assessing nutrient deficiency. More analyses are required to assess the robustness of this relationship, and also to assess if coarse-grained proteomic signatures are nutrient specific (i.e., would a coarse-grained marker be able to distinguish between Fe and Mn stress?).

CONCLUSION

We conclude that different microbial taxa have distinct coarse-grained proteomic composition, and this composition is more similar across taxa than across environmental conditions. The stoichiometry of proteins within pathways is conserved [9]—but our results show that this is not the case across pathways. Variation in pathway-to-pathway stoichiometry may indeed underpin ecological strategies, in addition to differing gene repertoires. Connecting in situ proteomes to ecological strategies will delineate proteomic traits, which can then be adopted into a trait-based approach for modeling microbial communities. Genomic trait-based approaches have successfully explained large-scale biogeochemical processes [10, 11], but they first had to identify genes that are metabolically important. Therefore, identifying and quantifying proteomic trait variation across taxa will connect protein production to ecological strategies, and ultimately enable modeling of microbial communities by representing proteomic traits and trade-offs in large-scale models (e.g., as in [70]).

DATA AVAILABILITY

The metagenomics and metatranscriptomics data reported here have been deposited in the NCBI sequence read archive (BioProject accession no. PRJNA074702; BioSample accession nos. SAMN18057468–SAMN18057479 (metagenomics) and BioSample accession nos. SAMN18057480–SAMN18057497 (metatranscriptomics). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository with the dataset identifier PXD022995 [71]. All other data products (the cobia analysis output, formatted databases, peptide abundances for each database search, targeted proteomics data, culture proteomics data, metaproteomic simulation output) are available in Dryad at https://doi.org/10.5061/dryad.vt4b8gtrz.

REFERENCES

1. Reimers AM, Knoop H, Bockmayr A, Steuer R. Cellular trade-offs and optimal resource allocation during cyanobacterial diurnal growth. PNAS. 2017;114:E6457–E6465.
2. Toseland A, Daines SJ, Clark JR, Kirkham A, Strauss J, Uhlig C, et al. The impact of temperature on marine phytoplankton resource allocation and metabolism. Nat Clim Change. 2013;3:1–6.
3. Twining BS, Baines SB. The trace metal composition of marine phytoplankton. Ann Rev Mar Sci. 2013;5:191–215.
4. Saito MA, Mcilvin MR, Moran DM, Goepfert TJ, Ditullio GR, Post AF, et al. Multiple nutrient stresses at intersecting Pacific Ocean biomes detected by protein biomarkers. Science. 2014;345:5–10.
5. Morris RM, Nunn BL, Frazar C, Goodlett DR, Ting YS, Rocap G. Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. ISME J. 2010;4:673–85.
6. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild *Prochlorococcus*. Science. 2014;344:416–20.
7. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. ISME J. 2014;8:1553–65.
8. Braakman R, Follows MJ, Chisholm SW. Metabolic evolution and the self-organization of ecosystems. PNAS. 2017;114:E3091–100.
9. Lalanne JB, Taggart JC, Guo MS, Herzel L, Schieler A, Li GW. Evolutionary convergence of pathway specific enzyme expression stoichiometry. Cell. 2018;173:749–61.
10. Reed DC, Algar CK, Huber JA, Dick GJ. Gene-centric approach to integrating environmental genomics and biogeochemical models. PNAS. 2014;111:1879–84.
11. Coles VJ, Stukel MR, Brooks MT, Burd A, Crump BC, Moran MA, et al. Ocean biogeochemistry modeled with emergent trait-based genomics. Science. 2017;1149:1–26.
12. Scott M, Gunderson CW, Mateescu EM, Zhang Z, Hwa T. Interdependence of cell growth and gene expression: origins and consequences. Science. 2010;30:1099–102.
13. Dekel E, Alon U. Optimality and evolutionary tuning of the expression level of a protein. Nature. 2005;436:588–92.
14. Parker DJ, Lalanne JB, Kimura S, Johnson GE, Waldor MK, Li GW. Growth-optimized Aminoacyl-tRNA synthetase levels prevent maximal tRNA charging. Cell Syst. 2020;11:121–30.
15. O'Malley MA, Parke EC. Microbes, mathematics, and models. Stud Hist Philos Sci A. 2018;72:1–10.
16. Johnson GE, Lalanne JB, Peters ML, Li GW. Functionally uncoupled transcription–translation in *Bacillus subtilis*. Nature. 2020;585:124–8.
17. Molenaar D, van Berlo R, de Ridder D, Teusink B. Shifts in growth strategies reflect tradeoffs in cellular economics. Mol Syst Bio. 2009;5:1–10.
18. Faizi M, Zavřel T, Loureiro C, Červený J, Steuer R. A model of optimal protein allocation during phototrophic growth. BioSystems. 2018;166:26–36.
19. Jahn M, Vialas V, Karlsen J, Ka L, Uhle M, Hudson EP. Growth of cyanobacteria is constrained by the abundance of light and carbon assimilation proteins. Cell Rep. 2018;25:478–86.
20. Hu SK, Liu Z, Alexander H, Campbell V, Connell PE, Dyhrman ST, et al. Shifting metabolic priorities among key protistan taxa within and below the euphotic zone. Environ Microbiol. 2018;20:2865–79.
21. Gifford SM, Sharma S, Booth M, Moran MA. Expression patterns reveal niche diversification in a marine microbial assemblage. ISME J. 2013;7:281–98.
22. Alexander H, Rouco M, Haley ST, Wilson ST, Karl DM, Dyhrman ST. Functional group-specific traits drive phytoplankton dynamics in the oligotrophic ocean. PNAS. 2015;112:E5972–9.
23. Alexander H, Jenkins BD, Rynearson TA, Dyhrman ST. Metatranscriptome analyses indicate resource partitioning between diatoms in the field. PNAS. 2015;112: E2182–90.
24. Sowell SM, Wilhelm LJ, Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, et al. Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. ISME J. 2009;3:93–105.
25. Russell JB, Cook GM. Energetics of bacterial growth: balance of anabolic and catabolic reactions. Microbiol Rev. 1995;59:1–15.
26. McGill BJ, Enquist BJ, Weiher E, Westoby M. Rebuilding community ecology from functional traits. Trends Ecol Evol. 2006;21:178–85.
27. Wu M, McCain JSP, Rowland E, Middag R, Sandgren M, Allen AE, et al. Manganese and iron deficiency in Southern Ocean *Phaeocystis antarctica* populations revealed through taxon-specific protein indicators. Nat Commun. 2019;10:3582.
28. Jabre L, Allen AE, McCain JSP, McCrow JP, Tenenbaum N, Spackeen JL, et al. Molecular underpinnings and biogeochemical consequences of enhanced diatom growth in a warming Southern Ocean. PNAS. 2021;118:e2107238118.
29. Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II global ocean sampling expedition: Northwest Atlantic through Eastern Tropical Pacific. PLoS Biol. 2007;5:0398–431.
30. Schmieder R, Lim YW, Edwards R. Identification and removal of ribosomal RNA sequences from metatranscriptomes. Bioinformatics. 2011;28:433–5.
31. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. Nucleic Acids Res. 2010;38:1–12.
32. Bertrand EM, McCrow JP, Moustafa A, Zheng H, McQuaid JB, Delmont TO, et al. Phytoplankton-bacterial interactions mediate micronutrient colimitation at the coastal Antarctic sea ice edge. PNAS. 2015;112:9938–43.
33. Podell S, Gaasterland T. DarkHorse: a method for genome-wide prediction of horizontal gene transfer. Genome Biol. 2007;8:R16.1–R16.18.
34. Enright AJ, Van Dongen S, Ouzounis CA. An efficient algorithm for large-scale detection of protein families. Nucleic Acids Res. 2002;30:1575–84.
35. Kanehisa M, Goto S. KEGG: Kyoto encyclopedia of genes and genomes. Nucleic Acids Res. 2000;28:27–30.
36. Kanehisa M, Sato Y, Kawashima M, Furumichi M, Tanabe M. KEGG as a reference resource for gene and protein annotation. Nucleic Acids Res. 2016;44:D457–62.
37. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, et al. The COG database: an updated version includes eukaryotes. BMC Bioinform. 2003;4:1–14.
38. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: the protein families database in 2021. Nucleic Acids Res. 2021;49: D412–9.
39. Haft DH, Selengut JD, White O. The TIGRFAMs database of protein families. Nucleic Acids Res. 2003;31:371–3.

40. McCain JSP, Tagliabue A, Susko E, Achterberg EP, Allen E, Bertrand EM. Cellular costs underpin micronutrient limitation in phytoplankton. Sci Adv. 2021;7:eabg6501.
41. Kim S, Pevzner PA. MS-GF+ makes progress towards a universal database search tool for proteomics. Nat Commun. 2014;5:1–10.
42. Rőst HL, Sachsenberg T, Aiche S, Bielow C, Weisser H, Aicheler F, et al. OpenMS: a flexible open-source software platform for mass spectrometry data analysis. Nat Methods. 2016;13:741–8.
43. Weisser H, Choudhary JS. Targeted feature detection for data-dependent shotgun proteomics. J Proteome Res. 2017;16:2964–74.
44. Weisser H, Nahnsen S, Grossmann J, Nilse L, Quandt A, Brauer H, et al. An automated pipeline for high-throughput label-free quantitative proteomics. J Proteome Res. 2013;12:1628–44.
45. Rőst HL, Schmitt U, Aebersold R, Malmstrőm L. pyOpenMS: a python-based interface to the OpenMS mass-spectrometry algorithm library. Proteomics. 2014;14:74–7.
46. Zybailov B, Mosley AL, Sardiu ME, Coleman MK, Florens L, Washburn MP. Statistical analysis of membrane proteome expression changes in Saccharomyces cerevisiae. J Proteome Res. 2006;5:2339–47.
47. Dupont CL, McCrow JP, Valas R, Moustafa A, Walworth N, Goodenough U, et al. Genomes and gene expression across light and productivity gradients in eastern subtropical pacific microbial communities. ISME J. 2015;9:1076–92.
48. McCain JSP, Bertrand EM. Prediction and consequences of cofragmentation in metaproteomics. J Proteome Res. 2019;18:3555–66.
49. Pfeifer N, Leinenbach A, Huber CG, Kohlbacher O. Statistical learning of peptide retention behavior in chromatographic separations: a new kernel-based approach for computational proteomics. BMC Bioinformatics. 2007;8:1–14.
50. Nunn BL, Faux JF, Hippmann AA, Maldonado MT, Harvey HR, Goodlett DR, et al. Diatom proteomics reveals unique acclimation strategies to mitigate Fe limitation. PLoS ONE. 2013;8:e75653.
51. Schmidt A, Kochanowski K, Vedelaar S, Ahrne E, Volkmer B, Callipo L, et al. The quantitative and condition-dependent Escherichia coli proteome. Nat Biotechnol. 2016;34:104–10.
52. Tanca A, Palomba A, Fraumene C, Pagnozzi D, Manghina V, Deligios M, et al. The impact of sequence database choice on metaproteomic results in gut microbiota studies. Microbiome. 2016;4:51.
53. Smith WO, Tozzi S, Long MC, Sedwick PN, Peloquin JA, Dunbar RB, et al. Spatial and temporal variations in variable fluorescence in the Ross Sea (Antarctica): oceanographic correlates and bloom dynamics. Deep Sea Res Part I Oceanogr Res Pap. 2013;79:141–55.
54. Andreoli C, Tolomio C, Moro I, Radice M, Moschin E, Bellato S. Diatoms and dinoflagellates in Terra Nova Bay (Ross Sea Antarctica) during austral summer 1990. Polar Biol. 1995;15:465–75.
55. Jeong HJ, Yoo YD, Kim JS, Seong KA, Kang NS, Kim TH. Growth, feeding and ecological roles of the mixotrophic and heterotrophic dinoflagellates in marine planktonic food webs. Ocean Sci J. 2010;45:65–91.
56. Sheftel H, Shoval O, Mayo A, Alon U. The geometry of the Pareto front in biological phenotype space. Ecol Evol. 2013;3:1471–83.
57. Hart Y, Sheftel H, Hausser J, Szekely P, Ben-Moshe NB, Korem Y, et al. Inferring biological tasks using Pareto analysis of high-dimensional data. Nat Methods. 2015;12:233–5.
58. Dethlefsen L, Schmidt TM. Performance of the translational apparatus varies with the ecological strategies of bacteria. J Bacteriol. 2007;189:3237–45.
59. Mori M, Schink S, Erickson DW, Gerland U, Hwa T. Quantifying the benefit of a proteome reserve in fluctuating environments. Nat Commun. 2017;8:1225.
60. Giovannoni SJ. SAR11 bacteria: the most abundant plankton in the oceans. Ann Rev Mar Sci. 2017;9:231–55.
61. Mangoni O, Saggiomo V, Bolinesi F, Margiotta F, Budillon G, Cotroneo Y, et al. Phytoplankton blooms during austral summer in the Ross Sea, Antarctica: driving factors and trophic implications. PLoS ONE. 2017;12:1–23.
62. Noble AE, Moran DM, Allen AE, Saito MA. Dissolved and particulate trace metal micronutrients under the McMurdo Sound seasonal sea ice: basal sea ice communities as a capacitor for iron. Front Chem. 2013;1:1–18.
63. Peloquin JA, Smith WO. Phytoplankton blooms in the Ross Sea, Antarctica: Interannual variability in magnitude, temporal patterns, and composition. J Geophys Res Oceans. 2007;112:1–12.
64. Sedwick PN, Garcia NS, Riseman SF, Marsay CM, DiTullio GR. Evidence for high iron requirements of colonial *Phaeocystis antarctica* at low irradiance. Biogeochemistry. 2007;83:83–97.
65. McQuaid JB, Kustka AB, Oborník M, Horák A, McCrow JP, Karas BJ, et al. Carbonate-sensitive phytotransferrin controls high-affinity iron uptake in diatoms. Nature. 2018;555:534–7.
66. Bertrand EM, Allen AE, Dupont CL, Norden-Krichmar TM, Bai J, Valas RE, et al. Influence of cobalamin scarcity on diatom molecular physiology and identification of a cobalamin acquisition protein. PNAS. 2012;109:E1762–E1771.
67. Cohen NR, Gong W, Moran DM, McIlvin MR, Saito MA, Marchetti A. Transcriptomic and proteomic responses of the oceanic diatom *Pseudo-nitzschia granii* to iron limitation. Environ Microbiol. 2018;20:3109–26.
68. Cohen NR, McIlvin MR, Moran DM, Held NA, Saunders JK, Hawco NJ, et al. Dinoflagellates alter their carbon and nutrient metabolic strategies across environmental gradients in the central Pacific Ocean. Nat Microbiol. 2021;6:173–86.
69. Strzepek RF, Harrisson PJ. Photosynthetic architecture differs in coastal and oceanic diatoms. Nature. 2004;431:689.
70. Follows MJ, Dutkiewicz S, Grant S, Chisholm SW. Emergent biogeography of microbial communities in a model ocean. Science. 2007;315:1843–6.
71. Perez-Riverol Y, Csordas A, Bai J, Bernal-Llinares M, Hewapathirana S, Kundu DJ, et al. The PRIDE database and related tools and resources in 2019: Improving support for quantification data. Nucleic Acids Res. 2019;47:D442–50.

ACKNOWLEDGEMENTS

We thank David Hutchins, Jeff Hoffman, Rachel Sipler, Jenna Spackeen, and Deborah Bronk, as well as Antarctic Support Contractors and the staff at McMurdo Station for support in the field. We are grateful to Alejandro Cohen from the Dalhousie Biological Mass Spectrometry Core Facility for assistance with mass spectrometry data acquisition, to Noor Youssef for discussion, and to Hong Zheng, John P. McCrow, and Pratap Venepally for contributions to sample purification, preparation for sequencing, and computational annotation. JSPM acknowledges support from the NSERC CREATE Transatlantic Ocean System Science and Technology Program and the Killam Scholarship. This project was financially supported by NSERC Discovery Grant RGPIN-2015-05009 to EMB, Simons Foundation Grant 504183 to EMB, an NSERC CGS Postgraduate scholarship to JSPM, NSF-ANT-1043671, NSF-OCE-1756884, and Gordon and Betty Moore Foundation Grant GBMF3828 to AEA.

AUTHOR CONTRIBUTIONS

JSPM, EMB, and AEA conceived the study. JSPM wrote the code, conducted analyses and metaproteomic lab work, with input from EMB and AEA. EMB and AEA collected the Antarctic samples and conducted metagenomic and metatranscriptomic lab and computational work. JSPM wrote the paper with input from EMB and AEA.

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/s41396-021-01084-9.

Correspondence and requests for materials should be addressed to Erin M. Bertrand.

Reprints and permission 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.