Inference of In Situ Microbial Physiologies via Sparse Tensor Decomposition of Metatranscriptomes

Application to Cyanobacteria Populations in the North Pacific

Stephen Blaskowski

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy

University of Washington
2024

Reading Committee:
E. Virginia Armbrust, Chair
Vaughn Iverson
William Stafford Noble

Program Authorized to Offer Degree:
Molecular Engineering and Sciences

University of Washington

Abstract

Inference of In Situ Microbial Physiologies via Sparse Tensor Decomposition of Metatranscriptomes: Application to Cyanobacteria Populations in the North Pacific

Stephen Blaskowski

Chair of the Supervisory Committee:
E. Virginia Armbrust
Department of Oceanography

Microbes respond to changes in their environment by adjusting their physiology through shifts in gene expression, which can be measured in the field by whole community RNA sequencing. The resulting metatranscriptomic data is inherently noisy, with unknown gene functions and fluctuations in organism abundance, all of which limit the utility of traditional methods. In the first chapter of this dissertation, I developed a novel statistical approach that uses sparse tensor decomposition to uncover patterns of gene co-expression. In the second chapter, I applied the method to metatranscriptomic data collected in the North Pacific, focusing on marine cyanobacteria, a group of highly abundant microbes that are responsible for up to a quarter of photosynthesis in global oceans. The analysis uncovered 25 robust co-expression patterns, including four that clarified how cyanobacteria respond in nature to scarce nitrogen and iron nutrients. In the final chapter I looked into another co-expression pattern that revealed how cyanobacteria respond to viral infection, placing this in the context of population diversity and evolution. Altogether this dissertation demonstrates the power of a new analytical approach to elucidate: 1) the functions of unknown genes, 2) how different organisms respond to environmental pressures, and 3) the ways in which microbial physiology and biogeochemical cycles interconnect in a changing ecosystem.

Acknowledgements

Like all scientific research, this work would have been impossible without an ecosystem of support, past and present, including intellectual contributions as well as those of a more personal nature. I want to extend my profound thanks to all these collaborators, and to specifically name some of those who have had a particular impact on my trajectory.

First and foremost to my coauthors – Marie Roald, Rogier Braakman, Paul Berube, Michael Carlson, and Debbie Lindell – it's been an unexpected joy to be influenced by the ways in which you view systems and problems. An extra thank you is due to Marie in particular. We have never met in person, and yet this research would not have existed without your guidance. What a marvel of the modern world to benefit from collaborations that span oceans, time zones, and cultures!

To the undergraduate researchers I've been privileged to work with – Jonah Valenti, Dhruvi Joshi, and Meena Shanmugam – your experiments, conversations, code, notes, and insights permeate this research, and our time working together has played a prominent role in my formation as a scientist. The paths you've embarked upon in your lives and careers fill me with optimism.

To my mentors, including Ginger Armbrust, Vaughn Iverson, and Bill Noble – thank you for your confidence in me when I couldn't find it in myself, your skepticism when I drifted towards astray, and your encouragement when my motivation flagged. Your nuggets of wisdom have echoed in my head as I've written this dissertation, and I'm grateful to carry them with me as I go. A special note of thanks to Ginger in particular, the best boss I've ever had, and a true master of the art of collaboration.

To my colleagues in the Armbrust and Noble labs, as well as colleagues in the Chisholm, Lindell, Ingalls, Bundy, Morris, and Rocap labs, and the researchers of the SCOPE and CBIOMES collaborations – your creativity, curiosity, and insight inspires me every day, and pushes me to be a better scientist.

To the scientists and crew of the Gradients expeditions – your tireless efforts collecting observations and samples on a rocky sea are the foundation of so much important research, including the work presented here.

To my beloved friends and family that lend purpose to this and every endeavor – there are too many to name everyone individually, but know that whether we have walked together since kindergarten or since last summer, whether by blood or by choice, you each hold a special place in my heart, and have uniquely motivated this work.

To Papa Henry Blaskowski – only recently has one of my papers started appearing ahead of your many patents when searching "Blaskowski" in Google Scholar. Thank you for passing the baton, and I hope I've done you proud!

To Dad – explorer, intellectual, and protector. Thank you for teaching me not to fear going off trail, and how to navigate when doing so.

To Mom – my greatest mentor and my first co-author! Thank you for walking slowly with me and teaching me how to wonder about this world.

To Sean – thank you for traversing the pandemic with me, for your wit, and for letting me practice my presentations with you.

To Evan – thank you for trusting me and holding my trust, for your earnestness, and for surviving parking lots and shepherd's tents with me.

To Grandma Katie McDonough – thank you for your friendship, your moral clarity, and for teaching me the importance of building and serving community. I remember you often, and carry your example with me always.

And ultimately, to Adrian – this work is as much a result of your sacrifice and dedication as it is mine. Words fail here, but nevertheless, thank you, and I love you.

Contents

List of Figures

List of Tables

0 Introduction

0.1 Motivation

Our planet's oceans have been the setting for several of the most pivotal episodes in the the ongoing ∼4 billion year saga of life on Earth. The leading theory on the origin of life, the RNA World hypothesis, proposes that the first biological systems were composed of self-replicating RNA macromolecules, the synthesis of which is thought to have been made possible under the conditions of an acidic primordial ocean [1]. This theory is supported by the remarkable biological versatility of RNA. As a single strand of nucleobases connected by a flexible ribose backbone, RNA can function both as a linear molecule encoding sequential, replicable genetic information, and as a three dimensional catalyst enabling the ebb and flow of metabolism [2]. Eventually, the more stable double stranded DNA came to supplant RNA as a specialized genetic molecule, and proteins offered a more flexible and effective suite of catalytic activity, but RNA retained its intermediary role in the Central Dogma of molecular biology, transcribing instructions from DNA and translating them into proteins. Presumably, this new complexity was established by the time of the emergence of the first cells, which, by 3.42 billion years ago, had colonized the shallow seas in the form of stromatolites and proto-microbial mats, according to fossil evidence [3]. About a billion years later, the oceans were again the setting for another singular event in Earth's history: the evolution of oxygenic photosynthesis in the microscopic ancestors of contemporary cyanobacteria [4]. The innovation of using water as the reductant for light-driven carbon fixation proved to be an evolutionary windfall. As far as we know, it has evolved only once in the history of life on Earth, and enabled oxygenic photoautotrophs to dominate the sunlit surfaces of the planet, leading to the rapid oxygenation of Earth's atmosphere and oceans [5]. This enduring change in the redox states of the air and sea paved the way for the evolution of more complex life. Eukaryotic phytoplankton carried forth the legacy of oxygenic photosynthesis by engulfing an ancestral cyanobacteria that would become the first plastid, eventually evolving into the chloroplasts that power plants and algae across both land and sea [6]. Despite a subsequent explosion in the diversity and complexity of biology enabled by these evolutionary events, the descendants of ancestral marine microbes continue to play an out-sized role in the dynamics of present day Earth systems [7]. This research was motivated by a desire to better understand these tiny, remarkable ocean inhabitants, and the molecular mechanisms by which they continue to drive and respond to planetary-scale systems.

The history of the study of microbial plankton is a story of technological advancement enabling increasingly deeper inquiry. The invention of the microscope led to the first observations of microscopic "animalcules" by inventor and naturalist Antoni van Leeuwenhoek [8]. Not long after, the first published drawing of a diatom initiated a wave of citizen science among wealthy European elites who were fascinated

by the symmetry and beauty of these microscopic aquatic organisms [9]. Centuries later, the discovery of *Prochlorococcus*, the smallest and most abundant photosynthetic organism on Earth, was made possible by the shipboard observation of seawater using flow cytometry, a technology developed for medical applications that was only 20 years old at the time [10]. Around the same time, the revolution in DNA sequencing was beginning to accelerate, eventually leading to the first published genomes of model marine microorganisms including the cyanobacterium *Prochlorococcus*, [11], the diatom *Thalassiosira pseudonana* [12], and *Pelagibacter ubique*, a representative of the cosmopolitan SAR11 clade of marine heterotrophic bacteria [13]. Further technological development enabled similar inventories of biological molecules beyond DNA, including transcriptome sequencing of RNA. In the years that followed, the advancement of massively parallel sequencing extended application beyond the minority of organisms cultured in a laboratory setting to whole communities of microorganisms sampled in their native environment. By simultaneously sequencing all of the transcripts in a sample, metatranscriptomics takes advantage of the central biologic role of RNA to allow us to peer into the identities, the evolutionary history, and the intermingled metabolism of the microbial communities teeming throughout the global oceans.

Technological innovation continues to be critical to furthering our understanding of marine microbes. Today, however, novel computational and statistical tools are as necessary as physical ones. The explosion of metatranscriptomic, metagenomic, and other metaomic data outpaces our capacity to interpret all the information these datasets contain [14]. The datasets are large and complex, and we need better means of extracting meaningful patterns. Many of the genetic sequences revealed by sequencing environmental samples have not previously been encountered by science, challenging existing methods of inferring biological function. Additionally, the tools for integrating metaomic datasets – for example combining metatranscriptomes with measurements of community genomes, proteins, and metabolites – are just beginning to be developed. In this research I aimed to expand the available toolbox for metatranscriptomic data analysis with the development of a novel methodology to help recover meaningful patterns from complex metatranscriptomic datasets. I then turned the tool's capacity towards understanding the in situ physiology of marine cyanobacteria. In concert with the countless insights and innovations of fellow researchers past and present, I hope these contributions will help elucidate the molecular comings and goings of our microbial cohabitants, and lend a deeper appreciation for how their histories and futures are intertwined with our own.

0.2 Challenges in metatranscriptomic analysis

Transcript sequencing data carries two linked information aspects: 1) the sequence of nucleotides in a particular RNA molecule and 2) the abundance of that RNA molecule. The RNA sequence can be used to

identify the organism of origin and assess the evolutionary history of the gene. The translated amino acid sequence contains information about the protein structure, which is intimately related to protein function. The abundance of an RNA transcript reflects the level to which the organism is expressing a particular gene at the moment of sampling. In the context of metatranscriptomics, the transcript abundance is a product of both the expression level of the gene, and the overall abundance of the organism in the community sample, the two of which can be difficult to disentangle. When examined together, the abundance levels of multiple genes across different samples can reveal sets of genes that are co-expressed, and thus may be responding to common environmental stimuli.

Analysis of metatranscriptomic data comes with significant algorithmic and computational challenges. Current metatranscriptomic datasets are mostly derived from massively parallel sequencing technologies, which output raw data in the form of short reads, each consisting of approximately 150 consecutive base pairs. These reads are random fragments of the originating sequences, which themselves constitute a mix of different transcripts from different organisms cohabiting a particular sample. Much like sifting through a bin of co-mingled shards of shattered ceramic vessels, computational algorithms must sift through and identify fragments that came from the same originating entity. This can be done in one of two ways. De novo assembly algorithms look for fragments with overlapping sequence similarity and piece them back together in an approximation of the original, much like looking for shards with contiguous patterns and gluing them back together. Mapping algorithms compare fragments back to a set of reference sequences assumed to be similar to the originals, much like pulling out shards that match an intact replica of the shattered vessel. Both methods come with considerable challenges, such as the computational power needed to run large assemblies [15], and the problem of identifying suitable reference sequences, especially when some of the sequences may not have been previously encountered [16].

Even once assembled or mapped, the analysis of processed metatranscriptomic sequence data is not straightforward. One challenge is how to infer the various functions of the transcripts found in a dataset. This is primarily done by comparing transcript sequences to databases of previously characterized genes, under the assumption that similar sequences are likely to encode similar functions. While existing annotation databases are extensive, and these homology-based techniques have proven to be powerful, there are drawbacks. Chief among these, functional inferences are limited by the existence of similar sequences. In contemporary metatranscriptomic and metagenomic samples, up to 50% of detected sequences share no significant homology with any of the references in common public databases; the so called "functional dark matter" of the global microbiome [17, 18]. Many analyses choose to simply discard these un-annotated sequences, despite the fact that some of them are widespread and abundant in datasets. These challenges highlight a need for alternative tools for inferring gene function in environmental sequencing data, and for

analytical methods that make use of transcript abundance data without the need for functional annotation.

Beyond functional annotation, metatranscriptomic datasets exhibit statistical properties that limit the relevance and utility of many existing computational and statistical tools [14]. Most environmental metatranscriptomes include hundreds of thousands if not millions of unique transcript sequences, but only tens to hundreds of unique samples, because of the technical and financial cost of sequencing each sample. Depending on the statistical task, this high dimensionality can result in under-powered analyses in which it can be difficult to extract useful inferences. This is compounded by the fact that metatranscriptomic data is typically noisy as a result of technical artifacts and biological stocasticity. Additionally, relatively few species and genes are ubiquitous in environmental samples, whereas there is often a long tail of rare species and genes that are sporadically detected in relatively few samples. This distribution results in sparse datasets filled with a high proportion of zero values, each of which can be interpreted differently as a unique function of organism abundance, gene expression, and the limit of detection in the sample. Addressing these statistical challenges is not straightforward, and many studies opt for tools originally developed for different data types, which though imperfect can offer reasonable solutions for specific tasks. As an alternative, in this work we developed a methodology that explicitly addresses the statistical properties typical of metatranscriptomic data, and in presenting it, I argue that such an approach enables more robust inference and more complete usage of the data.

0.3 From molecules to global cycles

Although the difference in scale between global ocean basins and biological molecules like RNA is many orders of magnitude, the collective activity of biological molecules accumulates to drive global scale processes. As noted above, the oxygen now abundant in Earth's oceans and atmosphere was the product of ancient cyanobacterial photosynthesis, and today their descendants continue to exercise an outsized influence on the redox state of the planet. Marine microbes account for roughly 50% of contemporary global primary production [19]. It's a remarkable statistic to cite, but the physical reality behind the statistic is perhaps even more remarkable. During photosynthesis, oxygen is produced at the site of the oxygen evolving complex, a collection of proteins that channel the light energy captured by chlorophyll, and use it to split a water molecule into oxygen and hydrogen, ultimately releasing an electron to be used in downstream biochemical reactions [20]. Despite decades of research, much of the fundamental biochemistry of this enzyme complex remains mysterious, including critical details around the reaction rate, mechanism, and intermediates [20, 21]. Nevertheless, during every hour of daylight, this reaction is repeated over and over, in thousands of photosystems per cell [22], in a number of cells that can only be estimated as many orders of magnitude greater than

the number of stars in the universe [23, 24]. It is through these nearly-unfathomable transformations and multitudes that the microscopic daily rhythms of microbial life mediate the flow of elements throughout the biosphere, and in doing so determine the composition of Earth's oceans, soils, and atmosphere.

In the current era of anthropogenic climate change, perhaps no biogeochemical cycle is as pertinent as the carbon cycle. Nearly all organic carbon is fixed via the catalytic activity of rubisco, another remarkable enzyme complex that leverages the energy produced by photosynthesis to convert inorganic carbon dioxide into sugar. Like the oxygen evolving complex, crucial details of the catalytic mechanism remain elusive, and efforts to engineer more efficient versions of the enzyme have so far proved unsuccessful [25]. Tracking with photosynthesis, total global carbon fixation is split nearly evenly between terrestrial and marine ecosystems. In contrast, the total carbon biomass on land is roughly 100 times that found in the oceans [26]. This discrepancy is the result of a rapid turnover of organic carbon in marine ecosystems, mediated largely by the microbial community. Viruses lyse up to 20% of microbial biomass per day, releasing dissolved organic carbon into the marine environment [27]. Most of this released organic matter is quickly consumed by heterotrophic microbes and respired as remineralized carbon dioxide, while recalcitrant dissolved organic carbon can remain in the water column for thousands of years [28]. This recalcitrant dissolved carbon, in combination with particulate organic carbon that sinks to the ocean depths, drives the marine "biological carbon pump" – one of several ways that Earth's oceans act as a net carbon sink [29]. Similar microbial networks mediate the flow of carbon through terrestrial soils [30], the oxidation and reduction of organic and inorganic nitrogen [31], the distribution of trace metals [32], and countless other elemental transformations essential to life. In each of these cycles, the equilibriums we observe at a planetary scare are dependent on the molecular details of the microbial physiologies involved.

Despite microbial metabolism being the engine that drives many global biogeochemical cycles, we are still in the early stages of incorporating microbes and their physiologies into models of elemental cycling. Efforts to model the interplay of microbial ecology with biogeochemistry are now several decades old [33], and yet parameter uncertainties still limit the predictive power of these models [34]. Many knowledge gaps contribute to these uncertainties, including incomplete inventories of the microbial actors, unknown ecological interactions, and incompletely characterized metabolic networks. Even the functions of well-studied biological processes are not fully understood, as illustrated by the mechanistic questions that remain around rubisco and the oxygen evolving complex. Continuing to characterize these central enzymes, in addition to the "functional dark matter" of currently uncharacterized microbial genes, is imperative to improve the utility and accuracy of biogeochemical models. In turn, improved modeling will help us predict the impacts of changing environmental conditions on organism physiologies and ecological systems, hopefully empowering humans to better navigate the compounding effects of climate change while holding in view the

impact on all of Earth's myriad forms of life.

0.4 Dissertation roadmap

This dissertation aims to contribute to advancing current understanding of the physiology and molecular biology of marine cyanobacteria, through the analysis of in situ metatranscriptomic sequencing data. In the first chapter, I detail development of a novel analytical tool to enable pattern discovery in complex metaomic datasets, with a particular focus on metatranscriptomics. In the second chapter, I describe deploying this tool on metatranscriptomes of marine cyanobacteria from the North Pacific. In doing so, I uncovered patterns of gene co-expression that offered insights into cyanobacterial physiology as it relates to iron and nitrogen nutrient limitation, and suggested hypothetical functions for previously uncharacterized genes. In the final chapter, I focus on another co-expression pattern that offered insights into the physiology, ecology, and evolution of cyanobacterial interactions with their phage predators. Collectively, this body of work offers a blueprint for a novel analytical approach that can help extract more robust insights from complex metaomic datasets. Finally, in the concluding remarks I comment on the ways in which microbial ecology and molecular engineering may help us address the growing challenges associated with climate change.

1 Development of a sparse tensor decomposition method for pattern discovery in metaomic datasets

This section contains content previously published as: Blaskowski S, Roald M, Berube PM, Braakman R, Armbrust EV. Simultaneous acclimation to nitrogen and iron scarcity in open ocean cyanobacteria revealed by sparse tensor decomposition of metatranscriptomes. bioRxiv. 2024:2024-07.

1.1 Introduction

In recent decades, metaomics technologies have proven to be powerful tools for studying microbial communities, generating comprehensive inventories of the DNA, RNA, proteins, and metabolites present in a microbiome [35]. Metatranscriptomic datasets are particularly information-rich because they catalog both the sequence of an RNA transcript, imprinted with the identity of the originating gene and organism, as well as the transcript's quantity, a composite reflection of organism abundance and gene expression. Thus, metatranscriptomes capture a snapshot of the in situ physiological dynamics of a microbiome. Commonly researchers infer these dynamics from metatranscriptomic data by comparing between sampling conditions the expression levels of genes involved in metabolic pathways of interest; however this approach is dependent on prior knowledge of the genes and pathways in question. Unsupervised clustering analysis provides an alternate avenue, independent of prior knowledge, which instead identifies groups of transcripts with similar abundance profiles across sampling conditions, aiming to uncover cohorts of co-expressed genes. These gene co-expression clusters reflect the modular organization of metabolic pathways and gene regulatory networks, and allow the functions of un-annotated genes to be inferred based on their association with better-characterized genes. Currently, the accumulation of metatranscriptomic data exceeds the advancement of analytical frameworks, and co-expression clustering represents a critical part of this gap.

Several co-expression clustering methods have been designed specifically for metatranscriptomic datasets [36, 37, 38], and additional studies have relied on methods originally designed for analyzing single-organism RNA sequencing data [39]. None of these approaches consider taxonomic information in the cluster detection algorithm, and resulting clusters can range from hundreds to thousands of genes, limiting interpretive power. Single-organism approaches are in turn ill-equipped to handle the idiosyncratic properties of metatranscriptomic data. Typical metatranscriptomic datasets are characterized by high dimensionality, variable community composition, overdispersion (variance increases exponentially with increasing mean), technical noise, and pervasive zero values, each presenting an obstacle to effective analysis [14]. And yet, the complexity of metatranscriptomic data also provides an opportunity. Metatranscriptomic datasets have a multiway

structure, in which each transcript abundance data point can be considered indexed by three variables: gene, taxon, and sample, with the latter serving as a proxy for the environmental context of expression. Analysis that effectively models this multiway structure can produce more robust inferences about the patterns of gene expression that give rise to metatranscriptomes, elucidating previously unrecognized microbiome dynamics.

Here we introduce Barnacle, a pattern discovery method developed to identify interpretable clusters of genes co-expressed across metatranscriptomes. Barnacle leverages the inherent multiway structure of metatranscriptomic data in combination with sparsity and non-negativity constraints to identify gene co-expression clusters that reflect the physiological states of organisms interacting in an environment. The foundation of the approach is CANDECOMP/PARAFAC (CP) tensor decomposition, a technique that models a multiway dataset as a sum of its constituent signals (for a review of tensor decomposition see Kolda and Bader, 2009 [40]). Component models, which include CP tensor decomposition, stand out among gene clustering techniques for their accuracy and robustness to noise, as well as their capacity to accommodate gene membership in more than one cluster, a property reflective of the structure of metabolic and regulatory networks [41]. Tensor decomposition methods have been developed for analyzing human gene expression datasets, including two studies that informed the formulation of model constraints used in Barnacle: the SDA algorithm that incorporates sparsity constraints using a Bayesian framework [42], and the MultiCluster method that employs partial non-negativity constraints [43]. Our approach represents a novel application of tensor decomposition to metaomic data that prioritizes the discovery of robust, interpretable inferences.

In this chapter we first describe the mathematical formulation of the sparse tensor decomposition model at the core of Barnacle. We then evaluate model performance using a dataset of 100 simulated data tensors that cover a range of shapes, sparsity levels, and noise levels. Next we introduce an innovative cross-validation procedure for parameter selection that leverages of sample replicates (a common feature of metaomic datasets). Finally, we evaluate the accuracy of this parameter selection procedure using the simulated data, and we assess the impact of sub-optimal parameter specification on model performance. The chapter concludes with a brief discussion of the utility of the method for analyzing metatranscriptomic data, in addition to the potential for application to other metaomics datasets.

1.2 Results

The sparse tensor decomposition model we developed is a constrained version of the classical CANDECOMP/PARAFAC (CP) tensor decomposition [44, 40]. We describe the model in the context of its intended application to environmental transcript abundance data. The notation and symbols used (Table 1.1) adhere to those laid out by Kolda and Bader (2009) [40]. Model implementation, documentation, and examples are

available as part of the Python package Barnacle.

Table 1.1: Main symbols used in the text.

Symbol	Definition
R	rank (number of components)
λ	sparsity coefficient
$\mathcal{Y}$	data tensor (gene × taxon × sample)
$\mathbf{G}$	gene component matrix
$\mathbf{T}$	taxon component matrix
$\mathbf{S}$	sample component matrix
$\mathbf{g}_r$	component r of matrix $\mathbf{G}$ (vector)
$\mathrm{Y}_{(\mathbf{G})}$	gene-mode unfolding of tensor $\mathcal{Y}$ (matrix)
$\|\cdot\|_F$	Frobenius norm
$\circ$	Outer product
$\odot$	Khatri-Rao product (column-wise Kronecker product)

1.2.1 Model

We consider a third-order tensor $\mathcal{Y}$ of transcript counts indexed in three modes: gene × taxon × sample. Each entry y_{ijk} encodes the abundance of gene i transcripts attributed to taxon j in sample k. The length of the gene mode, I, reflects the number of gene orthologs measured in the dataset; the length of the taxon mode, J, the number of taxa; and the length of the sample mode, K, the number of samples. The CP model represents the data tensor $\mathcal{Y}$ as a collection of R components (Fig 1.1A), each of which models a distinct linear signal pattern present in the data, obscured by a Gaussian noise tensor $\varepsilon \sim N(0, \sigma^2)$. That is

$$\mathcal{Y} \approx \sum_{r=1}^{R} \mathbf{g}_r \circ \mathbf{t}_r \circ \mathbf{s}_r + \varepsilon \tag{1.1}$$

or element-wise

$$y_{ijk} \approx \sum_{r=1}^{R} g_{ir} t_{jr} s_{kr} + \varepsilon_{ijk}, \tag{1.2}$$

where $\mathbf{G} \in \mathbb{R}^{I \times R}$, $\mathbf{T} \in \mathbb{R}^{J \times R}$, and $\mathbf{S} \in \mathbb{R}^{K \times R}$ are the component matrices with columns representing the weight vectors $\mathbf{g}_r$, $\mathbf{t}_r$, and $\mathbf{s}_r$ for the gene, taxon and sample modes, respectively. Entry g_{ir} indicates the relative contribution of gene i to the component r signal. When weight vector $\mathbf{g}_r$ is sparse, non-zero entries can be considered as a cluster of genes, unified by correlated expression profiles (Fig 1.1C). Weight vectors $\mathbf{t}_r$ and $\mathbf{s}_r$ indicate the relative activity of the component r signal in the modeled taxa and samples. Taking the outer product of weight vectors $\mathbf{g}_r$, $\mathbf{t}_r$, and $\mathbf{s}_r$ produces a rank-1 tensor representing component r (Fig 1.1B), and the sum of all R component outer products constitutes the rank-R tensor model, $\hat{\mathcal{Y}}$.

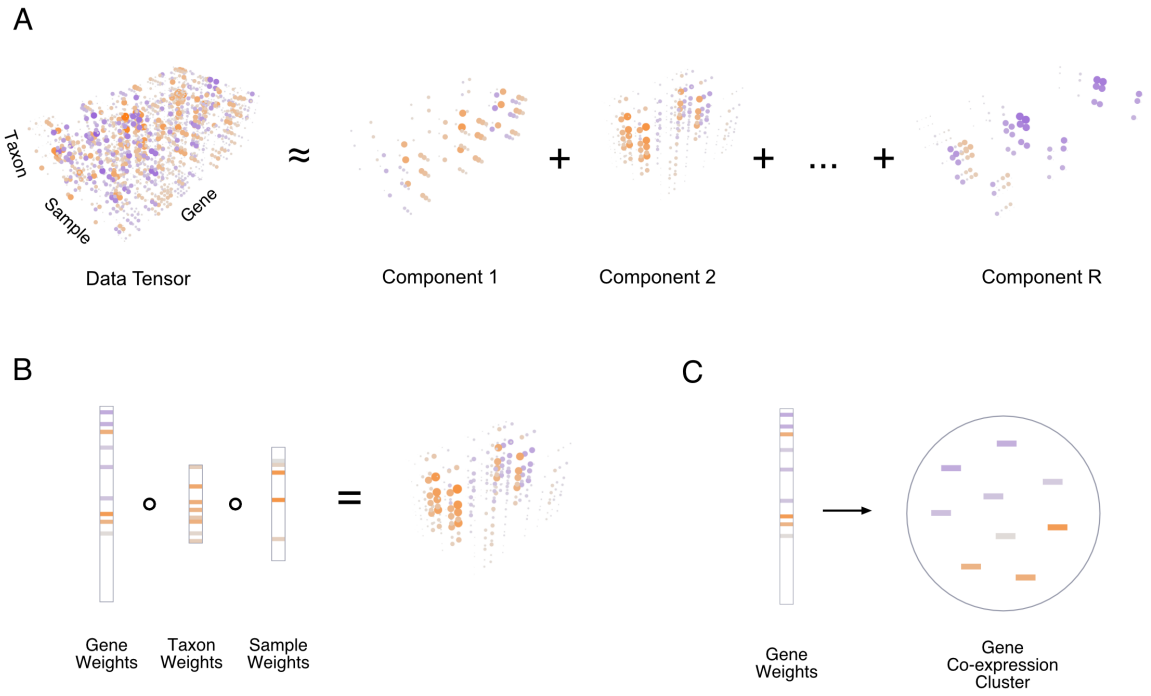

Figure 1.1: Diagram of sparse tensor decomposition model. (A) Visual representation of sparse CP tensor decomposition in which data tensor is modeled as a sum of R sparse components. Orange indicates positive values, purple negative values, and marker size indicates the magnitude of the value. (B) Diagram of a single component, which consists of gene, taxon, and sample weights. The outer product of these weight vectors equals a rank-1 tensor that represents the transcript abundance pattern modeled by the component. (C) Diagram illustrating the derivation of a gene co-expression cluster from the genes corresponding to non-zero weights in a component gene weight vector.

1.2.1.1 Constraints

Two sets of constraints are imposed on the component matrices of the tensor decomposition model to facilitate their interpretation. First, non-negativity constraints are imposed on the taxon and sample component matrices **T** and **S**. No non-negativity constraint is imposed on the gene component matrix **G**. This combination of non-negativity constraints abrogates issues of model indeterminacy resulting from sign flips while retaining the ability to model patterns of elevated and diminished expression (positively and negatively correlated transcript abundance profiles) within the same component. Second, we promote sparsity in the gene-mode component weight vectors via the incorporation of an l1 regularization penalty applied column-wise to the gene component matrix **G**. The degree of regularization is tuned by a sparsity coefficient parameter λ. This parameter indirectly controls the size of model clusters, in that a larger λ will result in more zero values in the gene component matrix **G**, and thus, on average, fewer genes corresponding to non-zero values in each component. The scaling indeterminacy of CP models (component matrices in

one mode can be arbitrarily scaled given an inverse scaling of the component matrices in the other modes) means that if norm-based regularization is imposed on at least one mode, then the component matrices of all modes must be regularized [45]. As such, the gene-mode sparsity penalty is accompanied by l2 unit norm constraints applied column-wise to the taxon and sample component matrices $\mathbf{T}$ and $\mathbf{S}$. Collectively, these constraints enable interpretation of components as clusters of genes with correlated transcript abundance patterns shared across the the taxa and samples indicated by component weights.

1.2.1.2 Optimization

The model is fit to data by minimizing a cost function (Eq (1.3)), parameterized by two user-defined variables: R, the number of components, and λ, the sparsity coefficient of the l1 regularization penalty applied to the gene-mode. The optimal solution to the cost function is obtained by component matrices $\mathbf{G}$, $\mathbf{T}$, and $\mathbf{S}$ that minimize the sum of square differences between the data tensor $\mathcal{Y}$ and the model tensor (the sum of R components), while also minimizing the sparsity penalty on the gene component matrix. Assuming the matrices fulfill Kruskal-rank conditions [46], this solution is guaranteed to be unique [44, 43].

$$\min_{\mathbf{G},\mathbf{T},\mathbf{S}} \sum_{ijk} \left(y_{ijk} - \sum_{r=1}^{R} g_{ir} t_{jr} s_{kr} \right)^2 + \lambda \sum_{r=1}^{R} \|\mathbf{g}_r\|_1 \text{ s.t. } \mathbf{T}, \mathbf{S} \geq 0; \|\mathbf{t}_r\|_2, \|\mathbf{s}_r\|_2 = 1 \tag{1.3}$$

We solve the cost function using a modified version of the alternating least squares (ALS) algorithm for CP tensor decomposition (see e.g. [40]). Briefly, the algorithm iterates over a series of update steps, each of which cycles between the modes of the tensor, updating the component matrix of one mode while the component matrices of the other two modes remain temporarily frozen. The component matrix of each mode is updated by solving a regularized least squares regression problem, calculated using the input data tensor (unfolded along the appropriate mode) along with the Khatri-Rao product of the two frozen component matrices (Eqs (1.4)-(1.6)). The updated component matrix entries are then frozen, and the algorithm proceeds to update the component matrix of the next mode. The algorithm iteratively updates each component matrix in this manner until the change in loss between successive iterations drops below a tolerance threshold.

$$\min_{\hat{\mathbf{G}}} \left\| \mathrm{Y}_{(\mathbf{G})} - \hat{\mathbf{G}}(\mathbf{T} \odot \mathbf{S}) \right\|_F^2 + \lambda \sum_{r=1}^{R} \|\hat{\mathbf{g}}_r\|_1 \tag{1.4}$$

$$\min_{\hat{\mathbf{T}}} \left\| \mathrm{Y}_{(\mathbf{T})} - \hat{\mathbf{T}}(\mathbf{G} \odot \mathbf{S}) \right\|_F^2 \text{ s.t. } \hat{\mathbf{T}} \geq 0; \left\| \hat{\mathbf{t}}_r \right\|_2 \leq 1 \tag{1.5}$$

$$\min_{\hat{\mathbf{S}}} \left\| \mathrm{Y}_{(\mathbf{S})} - \hat{\mathbf{S}}(\mathbf{G} \odot \mathbf{T}) \right\|_F^2 \text{ s.t. } \hat{\mathbf{S}} \geq 0; \|\hat{\mathbf{s}}_r\|_2 \leq 1 \tag{1.6}$$

Model constraints are accommodated by modifying the inner loop least squares subproblem for a given mode. First, updates of the **G** component matrix are formulated as a lasso problem to incorporate the l1 regularization (Eq (1.4)). Second, updates of the **T** and **S** component matrices are achieved using a least squares problem modified to include simultaneous enforcement of the non-negativity and l2 norm constraints (Eqs (1.5) and (1.6)). Note that to make these subproblems tractable, we relax the l2 norm constraint from strict equality to an inequality, which has the effect of unit equality of the l2 norms in the sample and taxon modes when combined with the l1 penalty in the gene mode. All inner loop subproblems are solved using the Fast Iterative Shrinkage/Thresholding Algorithm (FISTA) with constraints encoded as proximal operators [47]. To improve the convergence rate, we implemented the FISTA solver in combination with an adaptive restart scheme as proposed by O'Donoghue and Candès (2015) [48].

1.2.2 Model evaluation with simulated data

Model performance was evaluated using simulated data tensors, each constructed as the sum of a specified number of sparse components, combined with a Gaussian noise tensor. We measured performance on the basis of five metrics (Table 1.2): the sum of squared errors (SSE) to evaluate overall model fit, factor match score (FMS) to compare model components against ground truth components used to generate the simulation, and precision and recall metrics [41] to assess the similarity of model-derived gene sets to those derived from simulation components. Additionally, we calculated the harmonic mean of the precision and recall scores to produce the F1 score, providing a balanced measure of cluster accuracy.

The results of a representative simulation experiment (Fig 1.2) illustrate use of these metrics to evaluate model performance. In the experiment shown, we generated a simulated gene $\times$ taxon $\times$ sample tensor of shape $50 \times 20 \times 30$ using 8 components with an average gene cluster size of 20 (Fig 1.2D, left side) and a noise-to-signal ratio of 1. A series of models parameterized with different numbers of components (R) and sparsity coefficients (λ) were fit to the simulated data. Across a range of λ values, the model with the lowest SSE (comparing model to noiseless simulation) corresponded to $R = 8$ (Fig 1.2A), indicating that the most accurate model fit was achieved when the number of model components matched the true

Table 1.2: Metrics used to evaluate model performance.

Metric	Definition	Description
SSE	$\frac{\|\mathcal{Y}-\hat{\mathcal{Y}}\|_F^2}{\|\mathcal{Y}\|_F^2}$	Sum of Squared Error: Measure of how closely the model matches the data, scaled relative to the Froebenius norm of the data.
FMS	$\frac{1}{R}\sum_{r=1}^{R}\frac{\mathbf{g}_r^\intercal \hat{\mathbf{g}}_r}{\|\mathbf{g}_r\|\|\hat{\mathbf{g}}_r\|}$	Factor Match Score: Measure of how closely two component matrices match one another.
Precision	Saelens et al., 2018 [41]	Proportion of test cluster membership that adheres to the ground truth.
Recall	Saelens et al., 2018 [41]	Proportion of ground truth membership recapitulated by test clusters.
F1	$\frac{2(\text{precision})(\text{recall})}{\text{precision}+\text{recall}}$	Harmonic mean of precision and recall.

number of components used to generate the simulation. Given $R = 8$, we evaluated the effect of sparsity coefficient (λ) on model fit (SSE) and the accuracy of model components (FMS) and found that $\lambda = 0.8$ corresponded to both the minimum SSE and the maximum FMS (Fig 1.2B). However, the maximum F1 score, indicating the best accuracy of model-derived gene clusters, corresponded to $\lambda = 1.4$ (Fig 1.2C). This value of λ corresponded to the inflection points of the SSE and FMS curves rather than their respective minimum and maximum. The results also highlighted the inherent trade off between precision and recall: gene cluster recall improved at the expense of precision below $\lambda = 1.4$, and precision improved at the expense of recall above $\lambda = 1.4$ (Fig 1.2C). Importantly, the component matrices of the model parameterized with $R = 8$ (corresponding to minimum SSE), and $\lambda = 1.4$, (corresponding to maximum F1), closely resembled the ground truth components used to generate the simulation (Fig 1.2D).

Using these metrics, we assessed generalized model performance on 100 simulated tensors that collectively represented a range of tensor shapes, ranks, sparsity patterns, and noise-to-signal ratios ranging from 0.1 to 10 (Fig 1.3). The sparse tensor decomposition model was fit to each simulation using optimal parameters. The R parameter was set to the true number of components used to generate each simulation. The λ parameter does not correspond to any value used to generate simulations, so the optimal sparsity coefficient was estimated as the value of λ that yielded the maximum F1 score, which generally increased with the noise level and fraction of zero values of the simulation. Model fit, as measured by SSE, closely tracked simulated noise level (Fig 1.3A), indicating that the model represented the variation attributable to signal and did not over-fit to noise. Model components matched the component matrices used to generate simulations almost identically up to a noise-to-signal ratio of 1, above which FMS decreased to an average of about 0.6 at a noise-to-signal ratio of 10 (Fig 1.3B). The precision and recall of model-derived gene clusters were similarly high below a noise-to-signal ratio of 1 and exhibited a more gradual decline at higher noise levels compared to FMS, with average precision around 0.7 and average recall around 0.9 at a noise-to-signal ratio of 10 (Fig 1.3C and D). F1 scores reflected a balance between precision and recall with an average around 0.7 at

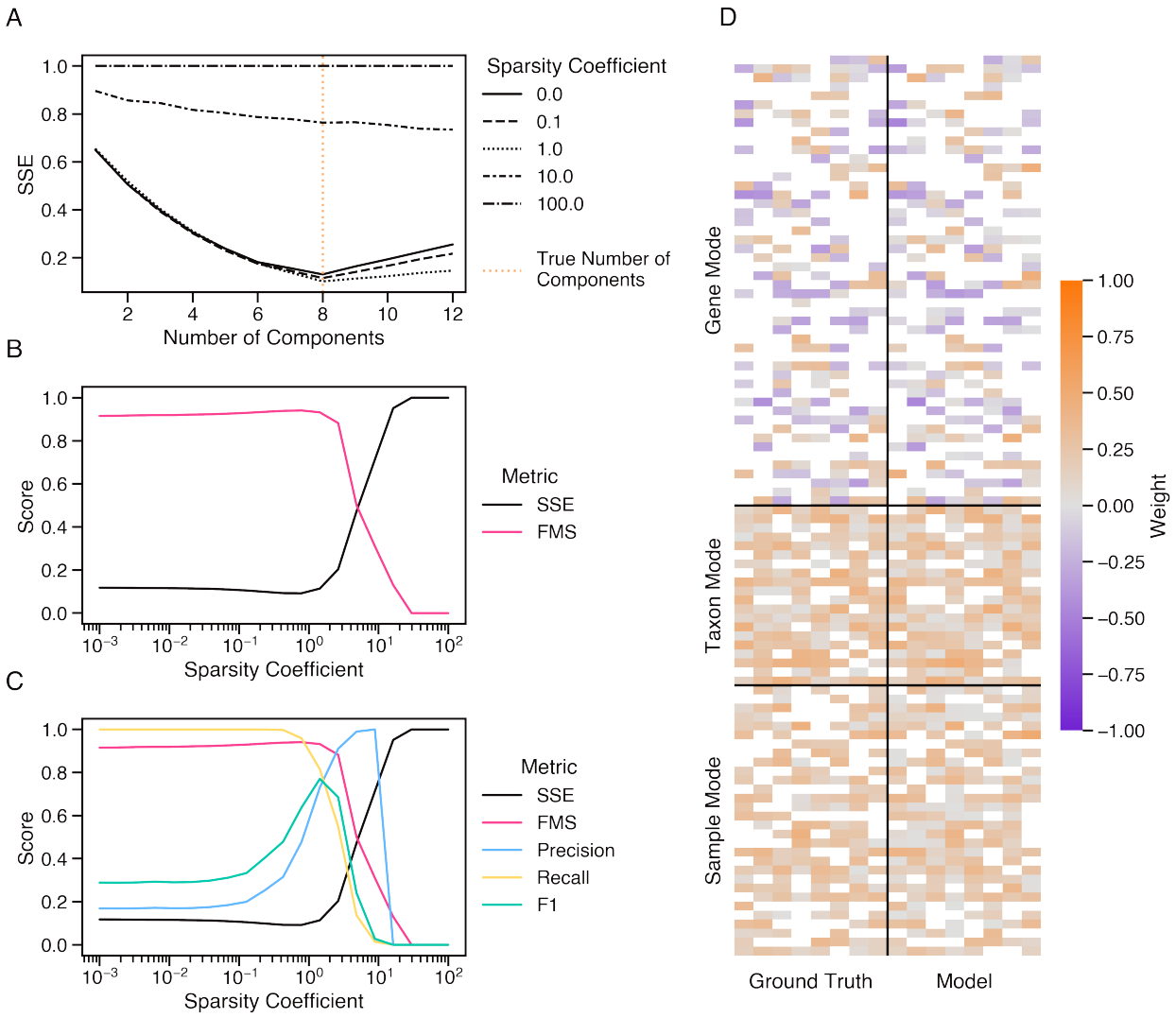

Figure 1.2: Example model evaluation with noisy simulated data tensor. (A) Changes in the relative sum of squared errors (SSE) resulting from parameterization with different numbers of components (R) and sparsity coefficients (λ). Vertical orange line indicates the number of components used to generate the simulation. Simulated noise level was scaled to a 1:1 noise-to-signal ratio. (B) Changes in the factor match score (FMS) and SSE resulting from parameterization with different values of λ (number of components fixed at $R = 8$). (C) Changes in the precision, recall, and F1 score of gene clusters derived from models parameterized with different λ values (fixed $R = 8$), plotted alongside SSE and FMS for ease of comparison. (D) Weights of a model parameterized with $R = 8$ and $\lambda = 1.4$ (right), shown in comparison to the components used to construct the simulation (left). Zero-valued weights are indicated by blank spaces.

the highest noise levels tested (Fig 1.3E). Overall these results indicate that the sparsity constraint is an effective counterbalance to noise, allowing the model to faithfully recover signal even in high noise datasets, and that this performance does not appear affected by rank, tensor shape, or the proportion of zero values.

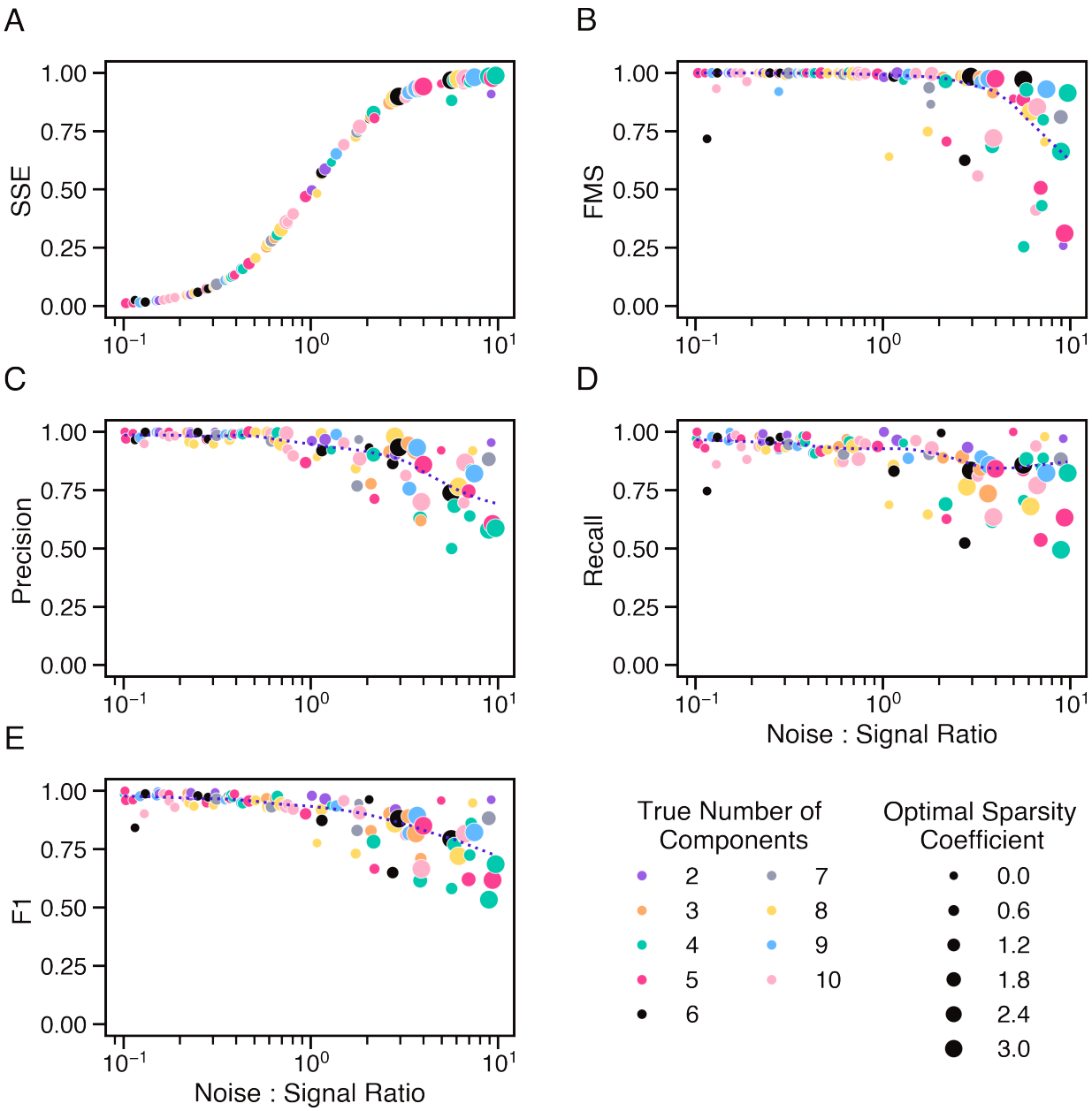

Figure 1.3: Generalized model performance on 100 noisy simulated data tensors. Performance of optimally-parameterized models of 100 simulated data tensors, evaluated in reference to simulation ground truth and plotted as a function of simulated noise-to-signal ratio. Optimal R parameter, indicated by marker color, was set equal to the true number of components used to generate each simulation, and optimal sparsity coefficient, indicated by marker size, was determined by maximum F1 score. (A) Overall model fit as measured by sum of squared errors (SSE). (B) Accuracy of component matrices as measured by factor match score (FMS). Accuracy of gene clusters resulting from model gene components, measured in terms of (C) precision, (D) recall, and (E) F1 score. Dotted lines indicate trends, as determined by locally weighted scatter plot smoothing (LOWESS) with a bandwidth of 0.3.

1.2.3 Parameter selection

Optimal model parameters are rarely known a priori and instead must be inferred from the data. We therefore developed a cross-validated grid search strategy that leverages sample replicates to select R and

λ parameters that result in the best fit model (see Methods). In brief, for each unique set of parameters, we fit a model to a subset of the data encompassing one replicate from each sampling condition, and then calculated cross-validated SSE and FMS scores using the held out replicates. We selected the R value of best fit based on the minimum cross-validated SSE. Fixing the R parameter, we then selected the best fit sparsity coefficient as the maximum λ value at which the cross-validated FMS remained within one standard error of the maximum FMS, a variation on the 1SE rule for parsimonious sparse model selection [49].

We assessed the effectiveness of the cross-validated grid search strategy using the 100 simulated tensors that collectively represented a range of tensor shapes, ranks, sparsity patterns, and noise-to-signal ratios ranging from 0.1 to 10. We simulated sample replicates by generating three identical copies of each signal tensor and combining each with an independent Gaussian noise tensor. We then independently fit models to each replicate tensor, calculated cross-validated SSE and FMS scores by comparing between replicates, and selected parameters of best fit. We selected the best fit R parameter as the number of components that resulted in the lowest cross-validated SSE, and among models fit with this value of R, we selected the best fit sparsity coefficient as the maximum λ value at which the cross-validated FMS fell within one standard error of the maximum FMS.

We compared R and λ parameters selected via cross-validated grid search with ground truth parameters, as determined by the true number of components used to generate each simulation and the value of λ that resulted in the maximum F1 score. The R identified by the grid search matched the true number of components in 86 of 100 simulations and was off by no more than 1 component in 92 simulations (Fig 1.4A). When the noise-to-signal ratio was greater than 1, the selected R in 13 simulations differed from the true number of components by an average of 2.2. At noise-to-signal ratios below this threshold, the selected R in a single simulation differed by one from the true number of components. The selected λ matched the optimal λ in 46 simulations and was within a twofold change in 80 simulations (Fig 1.4B). The frequency and magnitude of λ mis-specification was relatively consistent across simulation noise levels. Underestimation was more common than overestimation for both R and λ: 11 of the 14 R mis-specifications and 34 of the 54 λ mis-specifications undershot the ground truth parameter value. Taken together, these results demonstrated that the cross-validated grid search strategy selects parameters that approach optimal values, even up to a noise level ten times that of the signal.

1.2.4 Robustness to mis-specification

We also examined the ramifications of inaccurate parameter selection in models fit to the 100 simulated data tensors. We aligned all models against the ground truth components used to generate each simulation, and assessed SSE, FMS, precision, recall, and F1 scores in reference to a noiseless version of the simulated

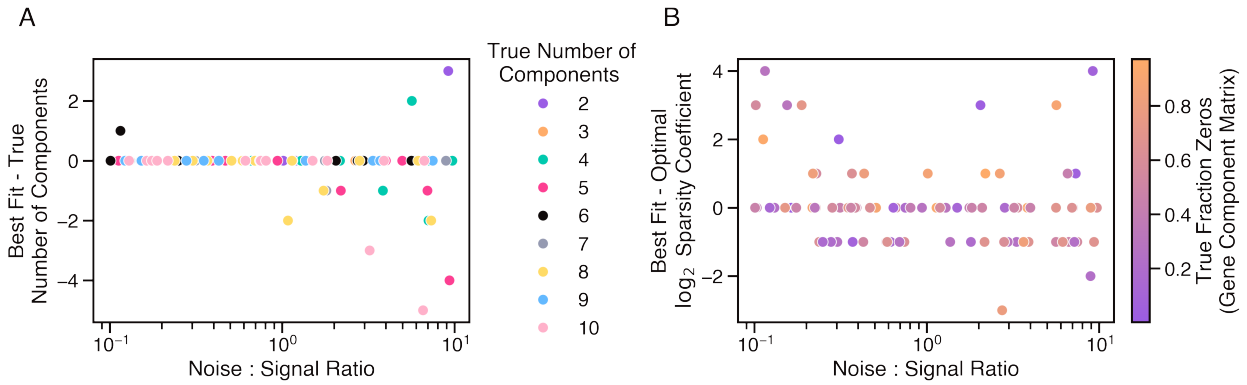

Figure 1.4: Cross-validated grid search identifies best fit parameters that approach ground truth in 100 simulated datasets. (A) Difference between model R parameter (number of components) selected via cross-validated grid search, and the true number of components in each simulation, plotted against simulation noise level and colored by true R. (B) Log-2 change between λ parameter (sparsity coefficient) selected via cross-validated grid search, and the optimal λ that maximizes the mean F1 score between model-derived gene clusters and those of the simulation ground truth. Scores are plotted against simulation noise level and colored by the true fraction of zero values in the simulation gene component matrix. Six data points in which the optimal or best fit λ was 0 were excluded for ease of visualization.

dataset. In most cases, the FMS, precision, recall, and F1 scores showed little difference between models with mis-specified or true R parameter values (Fig 1.5A-D). In all but one case, the F1 score of models parameterized with an inaccurate R was nearly identical to the optimally-parameterized model, highlighting that model-derived clusters are robust to R mis-specification. When λ was underestimated, the resulting gene clusters generally exhibited lower precision and higher recall, whereas when λ was overestimated, the clusters exhibited higher precision and lower recall (Fig 1.5E-H). These data suggested that although high-sparsity models may incorrectly exclude some genes from clusters, increasing sparsity provides greater confidence in the retained composition of modules, that they accurately reflect true cluster structure in the underlying data.

1.3 Discussion

This chapter details the development and evaluation of Barnacle: a novel tool for unsupervised pattern discovery in metaomic data. In particular, our design was motivated by a need for analytical tools suitable for detecting co-expression clustering patters in environmental metatranscriptomic datasets. Using simulated data, we verified that Barnacle faithfully recovers overlapping signals from noisy datasets (Fig 1.2), and found that this performance generalizes well across a range of tensor shapes, sparsity levels, and noise levels (Fig 1.3). We also developed and evaluated a cross-validation strategy that leverages sample replicates to identify model parameters that result in the best model fit. Experiments with the simulated data demonstrated the reliability of this parameter selection strategy (Fig 1.4), and indicated model robustness in cases

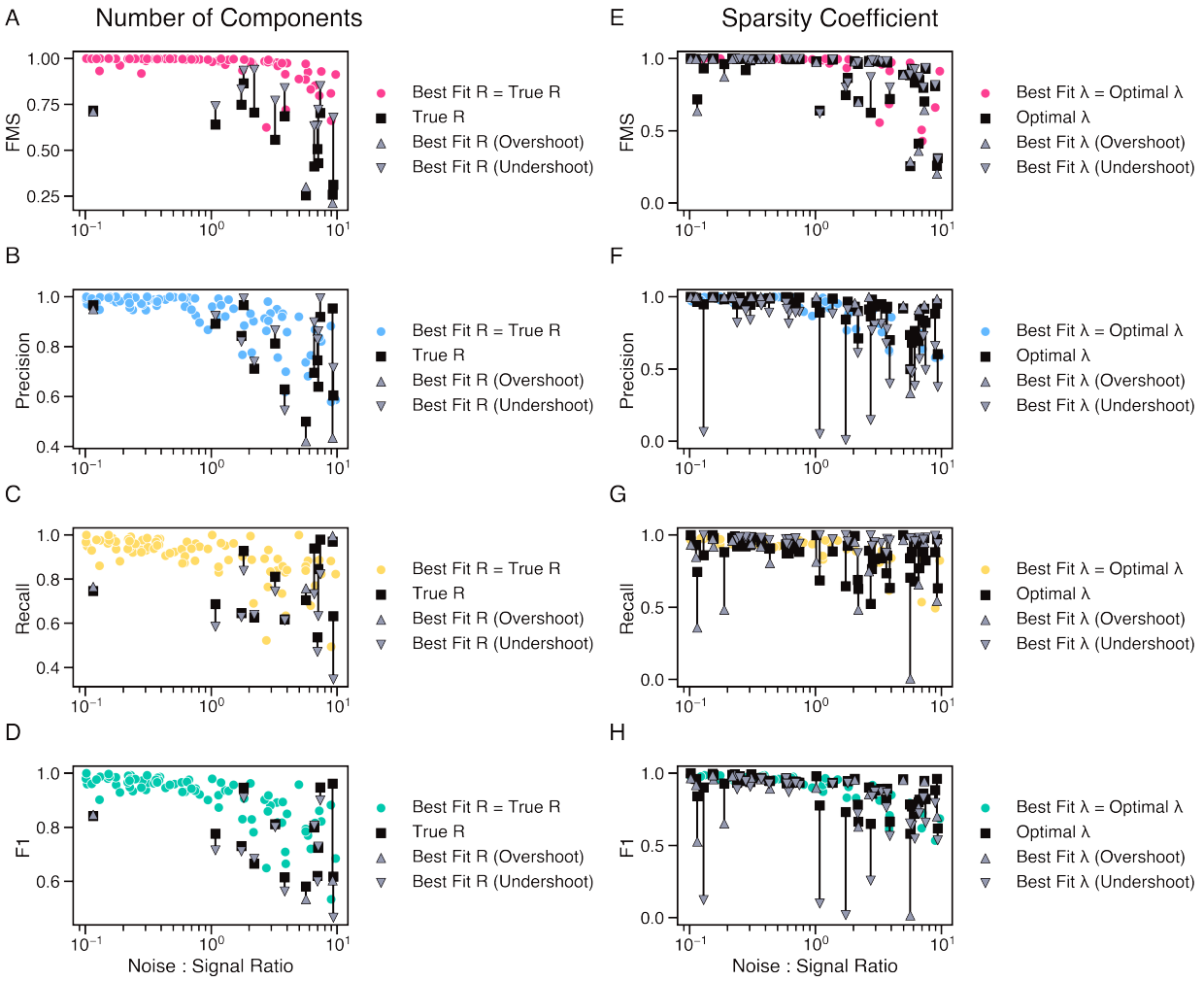

Figure 1.5: Evaluation of effect of parameter mis-specification on model component accuracy in 100 simulated datasets. The effect of R parameter (number of components) mis-specification on (A) overall model factor match score (FMS), (B) gene cluster precision, (C) recall, and (D) F1 score, evaluated in comparison to simulation ground truth. The effect of λ parameter (sparsity coefficient) mis-specification on (E) overall model FMS, (F) gene cluster precision, (G) recall, and (H) F1 score, evaluated in comparison to simulation ground truth. Colored points indicate simulations in which cross-validated grid search identified the optimal parameter. In cases where cross-validation undershot or overshot the optimal parameter, black bars indicate the difference in score between model fit with mis-specified parameter (grey triangles) and optimal parameter (black square).

where parameters were mis-specified (Fig 1.5). Altogether, the results demonstrated the functionality of a novel tool that promises to uncover from metaomic datasets previously obscured insights about the structure and ecology of microbiomes.

In the design of Barnacle we aimed to balance modeling the complexity of microbiome dynamics with incorporating constraints that promote intuitive interpretation. The core of the method, tensor decomposition, reflects the inherent structure of metaomic datasets. For example, metatranscriptomics datasets quantify gene expression, with each transcript abundance value indexed by the gene, taxon, and sample of origin

(Fig 1.1A). Metagenomes and metaproteomes have an analogous structure, with tensor values filled with copy number counts and protein abundances, respectively, instead of transcript abundances. Performing the decomposition on multiway tensor data, as opposed to two-way matrix decomposition, also serves as a fundamental model constraint [44], focusing Barnacle on the patterns that arise from linear combinations of genes, taxa, and samples. We further constrained the decomposition problem by imposing non-negativity and sparsity constraints on model components, focusing analysis on the most dominant patterns and minimizing false positives in the resulting gene sets. In addition to lending interpretability to the model, these constraints are a bulwark against noise, as illustrated by our experiments that demonstrated strong component recovery and a controlled false discovery rate even at a noise-to-signal ratio of 10:1 (Fig 1.3). Collectively these model design choices function to lend increased confidence to the inferences and associations derived from metaomic analysis.

Although large metaomics datasets are now common in studies of environmental microbiomes, the biological insights contained in these datasets often remain obscured by data complexity that exceeds the capacity of traditional analytical techniques. Barnacle was designed to help bridge this technological gap. The sparsity and non-negativity constraints make Barnacle well-suited to the high numbers of zero values and noise typical of environmental metaomic datasets. The unsupervised nature of our method alleviates dependency on functional annotations, enabling the concordant analysis of novel, uncharacterized gene families. This capability is critical as new surveys of environmental gene sequences routinely uncover a majority with no resemblance to previously described gene families [16, 50, 51, 18]. Barnacle offers a window of insight into the roles of these unknowns via their association with better characterized genes. In the future, this association information could be combined with data on gene co-evolution, structural modeling, and synteny to bolster computationally-derived functional inferences. Finally, our approach is flexible, and could be applied to any multiway dataset that can be represented as a three-way tensor, including metagenomics, proteomics, and metabolomics. In these ways, the development of Barnacle offers a novel unsupervised analytical technique that will better equip microbiome researchers to disentangle complex datasets, and to discern the patterns by which molecules, cells, and ecosystems interact to drive global biogeochemical cycles.

1.4 Methods

The code and data necessary to reproduce the analyses and evaluate the conclusions in the chapter are available as part of the publication Blaskowski et al. (2024) [52]. Supplementary Data files are available in an associated Zenodo repository: https://doi.org/10.5281/zenodo.12210994. Source code and documentation for the Barnacle model can be found in the package repository: https://github.com/blasks/barnacle. All

code required to process data, run analyses and produce figures presented in this manuscript can be found in an associated manuscript repository: https://github.com/blasks/barnacle-manuscript.

1.4.1 Implementation

Our implementation of the sparse tensor decomposition model presented here is freely available as the Python package Barnacle. Much of Barnacle's core functionality is built on top of Tensorly [53]. The model itself can be accessed via the 'SparseCP' class, modeled after Tensorly's standardized decomposition API. Our implementation of the FISTA algorithm with adaptive restart [47, 48] for solving constrained inner loop least squares problems is under the 'fista' module of Barnacle. Code for constructing and manipulating the simulated data tensors used in model development and evaluation is available under the 'tensors' module. Assorted functions including visualization tools and our implementation of the cluster evaluation scores developed by Saelens et al. (2018) [41] are collected under the 'utils' module.

1.4.2 Convergence

Convergence was defined using loss, calculated as

$$l^{(n)} = \sum_{ijk} \left(y_{ijk} - \sum_{r=1}^{R} g_{ir} t_{jr} s_{kr} \right)^2 + \lambda \sum_{r=1}^{R} \|\mathbf{g}_r\|_1 \tag{1.7}$$

where $l^{(n)}$ designates the loss at iteration n of the modified ALS algorithm. In all experiments performed in this study, we considered the algorithm to have converged when the change in loss dropped below 10^{-5}, or equivalently, when the following inequality was satisfied:

$$l^{(n-1)} - l^{(n)} < 10^{-5} \tag{1.8}$$

This change in loss is always non-negative because the loss decreases monotonically with each iteration of the ALS algorithm.

Depending on the initialization, it is possible for the ALS algorithm to converge on a local minimum rather than the globally optimal solution [40]. To mitigate this issue, we repeated all decompositions with five different random initializations, and the solution corresponding to the lowest loss was saved for analysis. Convergence was verified for each initialization based on the loss criterion outlined (Eqs (1.7) and (1.8)); all decompositions performed in this study converged in under 2,000 iterations. The random state of each decomposition was initialized with a unique integer seed, and to ensure reproducibility the seed of each saved model was stored as a local text file alongside model solutions and parameters.

1.4.3 Simulated Data

Development and evaluation of the sparse tensor decomposition model relied on simulated data. All simulations were third-order tensors constructed as the outer product of three randomly generated component matrices, constituting a signal tensor, in combination with a randomly generated Gaussian noise tensor. We varied the tensor shape, number of components, sparsity of the component matrices, and noise-to-signal ratio of each simulation depending on the experiment. Component matrix entries were randomly drawn from uniform distributions parameterized according to mode: $U(-1, 1)$ for the simulated gene mode to allow for negative and positive values, and $U(0, 1)$ for the simulated taxon and sample modes. The length of the simulated gene mode was randomly drawn from the interval $(10, 1000)$. The lengths of the simulated taxon and sample modes were independently and randomly drawn from the interval $(10, 100)$. The number of components (R) was randomly drawn from the interval $(2, 10)$. The fraction of zero values in each component matrix was independently stipulated by drawing a value from a uniform distribution $U(0, m)$ where $0 \leq m \leq 1$ represents the maximum number of zeros possible so that the matrix remained full rank. Noise tensor entries were randomly drawn from a normal distribution $N(0, 1)$, and then scaled them in proportion to the magnitude of the signal tensor. For each simulation, we specified the noise-to-signal ratio with a scaling factor s, where $s = 10^x$ with x being randomly drawn from a uniform distribution $U(-1, 1)$. To maintain sparsity in the final simulated data tensor, we set to zero each element of the noise tensor that corresponded to a co-localized zero value in the signal tensor.

Parameter selection

To select appropriate values of R and λ, we developed a cross-validated grid search strategy that made use of sample replicates common in metaomics datasets. With real data, tensors would be split by replicate along the sample axis to produce three replicate subtensors of shape $I \times J \times K_A$, $I \times J \times K_B$ and $I \times J \times K_C$, where I is the number of genes in the full dataset, J is the number of taxa in the full dataset, and K_A, K_B and K_C are the number of samples in replicate set A, B and C, respectively. For experiments with simulated data, we simulated replicates by generating three independent noise tensors, and combining each with the same underlying signal tensor. For both the simulation and real data tensors, we fit a series of models to each replicate subtensor using a grid search of different R and λ parameter values. Six cross-validated SSE scores were calculated for each unique set of parameters by comparing each fit model against the two held out replicate subtensors. Three cross-validated FMS scores were calculated for each parameter set by comparing the components between each pair of replicate models. In the simulated data experiments, every combination of parameters $R \in [1, 2, 3, ..., 12]$ and $\lambda \in [0.0, 0.05, 0.1, 0.2, 0.4, 0.8, 1.6, 3.2, 6.4, 12.8]$ was tested

against each simulation. We examined the cross-validated SSE scores of the $\lambda = 0.0$ models to identify the the minimum error model in the absence of l1 regularization. The number of components was set to the R value corresponding to the this minimum. We then selected the sparsity coefficient as the maximum λ value at which the cross-validated FMS remained within one standard error of the maximum FMS.

2 Simultaneous acclimation to nitrogen and iron scarcity in open ocean cyanobacteria

This section contains content previously published as: Blaskowski S, Roald M, Berube PM, Braakman R, Armbrust EV. Simultaneous acclimation to nitrogen and iron scarcity in open ocean cyanobacteria revealed by sparse tensor decomposition of metatranscriptomes. bioRxiv. 2024:2024-07.

2.1 Introduction

Microbial communities play essential roles in every Earth biome and drive global biogeochemical cycles through their collective metabolic activity [54]. In open ocean ecosystems, the abundant cyanobacteria *Prochlorococcus* and *Synechococcus* are keystone photosynthetic microbes that account for an estimated $\sim 20-25\%$ of total marine primary production [55, 56]. Although descended from a common ancestor, these two genera encompass a diversity of genetically and phenotypically differentiated subpopulations [57]. *Prochlorococcus* numerically dominates in the nutrient-poor tropical and subtropical open ocean waters that span from about 40°S to 40°N, whereas *Synechococcus* thrives across a greater range of habitats and is more abundant in colder, more nutrient-rich regions [55, 56]. Within each genus, genetically differentiated clades are further adapted to occupy distinct ecological niches, such as high-light adapted *Prochlorococcus* clades encountered near the surface, and low-light adapted *Prochlorococcus* clades generally found in deeper waters [23]. This ecological partitioning is driven by the functional diversity of an extensive pangenome [58]. *Prochlorococcus* and *Synechococcus* genomes share a core of about 1,000 – 1,500 genes, and a suite of accessory genes enable different clades and strains to adapt to different environmental conditions [59, 23]. Cultured isolates do not yet represent the full breadth of this sequence diversity, and a majority of the cyanobacterial pangenome remains functionally uncharacterized – part of the "functional dark matter" that constitutes a majority of gene families in the overall global microbiome [18]. Moreover, metabolism emerges not from the activity of independent genes, but rather from multiple gene products acting in concert. Characterizing these functional gene networks in marine cyanobacteria is crucial for clarifying the metabolic links between microbial genes and the ecosystem-scale processes they drive.

In the preceding chapter we described the construction and validation of Barnacle, a sparse tensor decomposition model for exploratory pattern discovery in metaomics datasets. In this chapter we turn Barnacle's capabilities to the analysis of 222 marine metatranscriptomes collected during three cruises in the North Pacific Ocean, which transited along the 158th meridian west (Fig 2.1), sampling from the nitrogen-limited subtropical gyre to the southernmost edge of the iron-limited subarctic gyre [60]. The transition zone be-

tween the two gyres is a dynamic interface in which elevated phytoplankton productivity supports a high diversity of marine life, including albacore tuna and loggerhead turtles that follow the front as a migratory route [61]. We focused our analysis on the cyanobacterial communities that form the foundation of these ecosystems and uncovered gene clusters that underlie acclimation of *Prochlorococcus* and *Synechococcus* sub-populations to shifting environmental pressures. This work demonstrates the power of unsupervised signal discovery techniques to draw from metatranscriptomics data explanatory connections between the molecular functions of individual genes, the cellular programs these genes drive, and the ecosystem scale processes that emerge from the collective metabolic activities of interacting microbes.

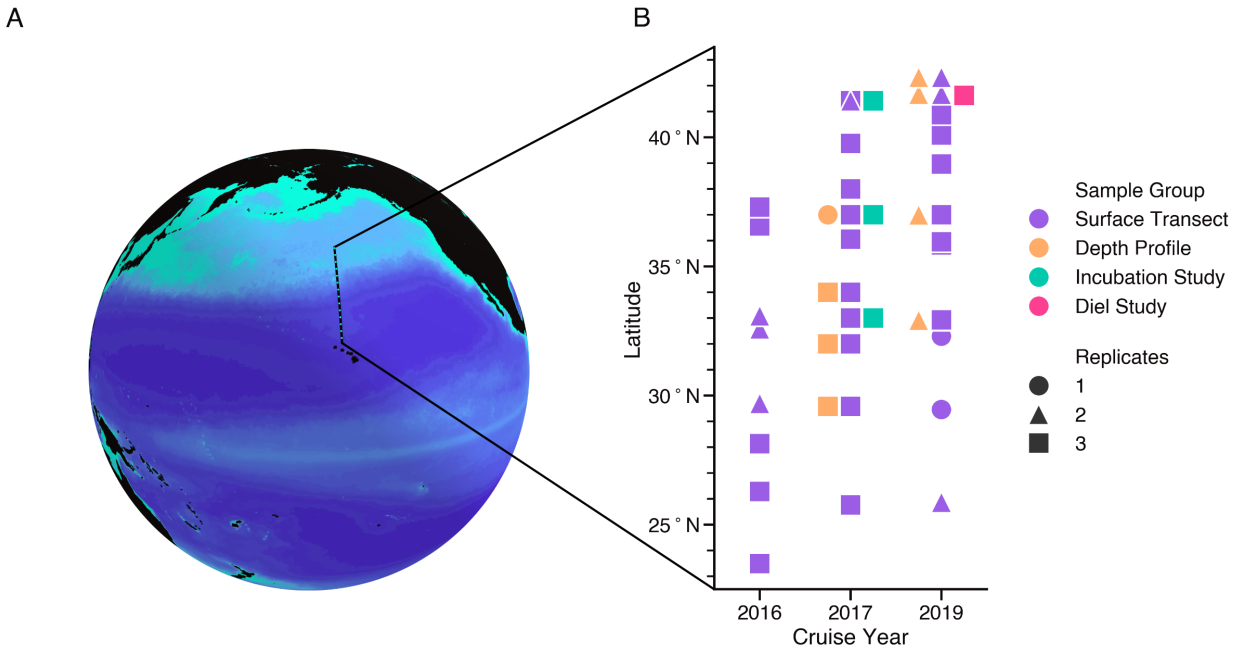

Figure 2.1: Summary of sampling locations and datasets integrated in this study. (A) Approximate cruise track (dotted line) along 158th meridian west, plotted over April climatology of surface chlorophyll, as measured by the MODIS Aqua satellite. (B) Sampling latitudes by cruise year, colored by dataset. Marker shape indicates maximum number of replicates retrieved for each sample set. Each marker for a depth profile, incubation study, or diel study encompasses multiple samples taken at the same latitude at different depths, treatments, and times of day, respectively.

2.2 Results

In an effort to identify co-expression patterns driving open ocean cyanobacterial ecosystem dynamics, we sought to apply the sparse tensor decomposition model to *Prochlorococcus* and *Synechococcus* community gene expression data. To this end, we compiled a dataset of 222 metatranscriptomes collected across different years and locations in the North Pacific Ocean (Fig 2.1). We mapped the metatranscriptome sequencing reads against a database of 681 *Prochlorococcus* and *Synechococcus* reference genomes [62] to determine the

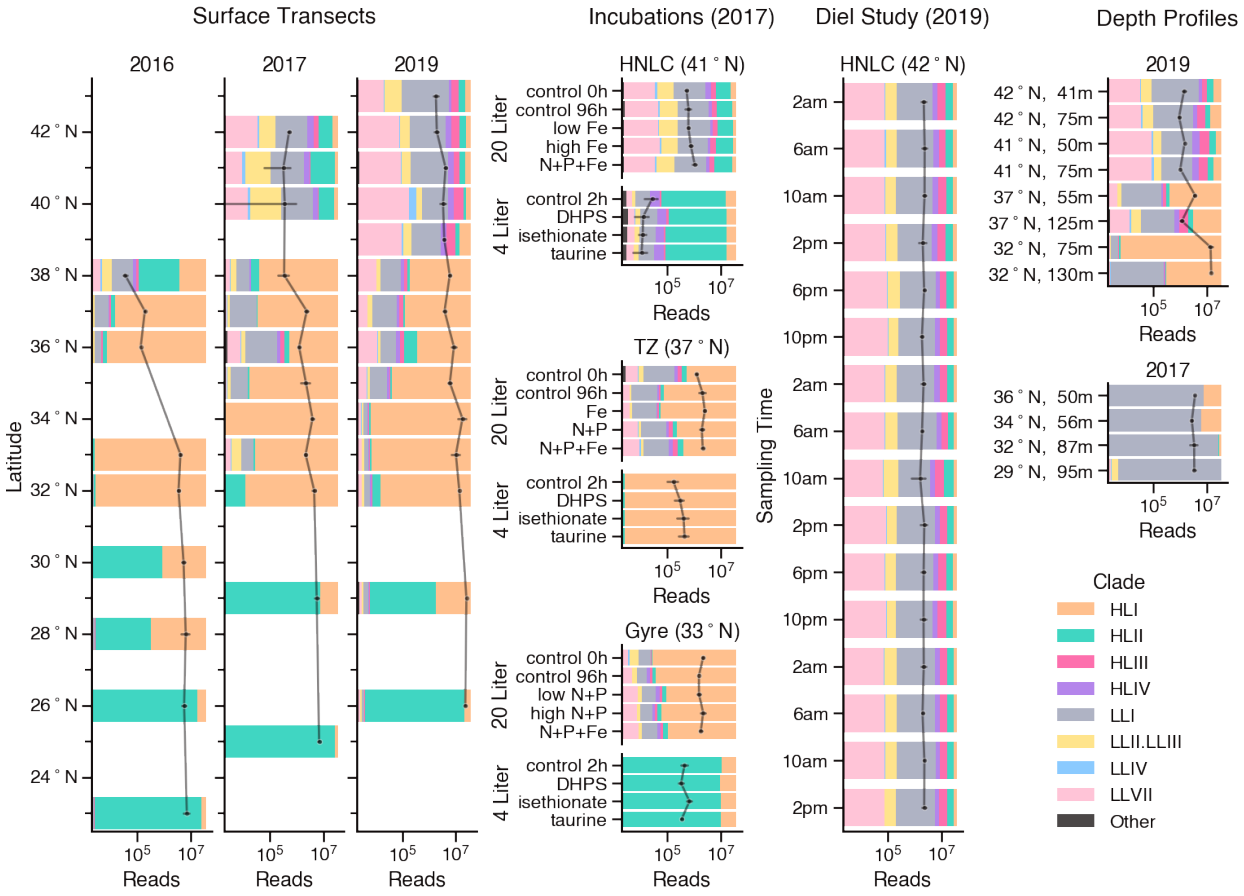

Figure 2.2: Composition of *Prochlorococcus* community transcript sequencing reads. Relative abundance of transcript reads mapped to *Prochlorococcus* clades. The size of each colored region is proportional to the fraction of sample reads mapped to the indicated clade (sample replicates averaged) with clade fractions summing to one in each sample, and samples grouped by dataset. X-axis corresponds to black line, showing aggregate count of *Prochlorococcus* transcript reads per sample with error bars indicating standard deviation of replicates.

abundances of transcripts originating from these species. Sequencing read counts were then aggregated by Cyanobacterial Clusters of Orthologous Groups of proteins (CyCOGs) and taxonomic clade. Clade aggregated read counts revealed a latitudinal shift in cyanobacterial community composition that was relatively consistent between cruise years. The *Prochlorococcus* community was dominated by the HLII clade within the North Pacific Subtropical Gyre (NPSG), shifted to HLI dominance in the transition zone, and gave way to a smaller community primarily consisting of LLI and LLVII clades near the southern subarctic gyre (Fig 2.2). While the *Synechococcus* community showed more variability, predominate trends revealed the dominance of clade II in the NPSG, a rise in the community share of CRD1 and CRD2 in the transition zone, and the prevalence of clades I and IV near the subarctic gyre, as well as some detection of subclusters 5.2 and 5.3 at the northernmost extent of the transects (Fig 2.3). These observed shifts in community composition underscore the dynamic nature of North Pacific ecosystems, and highlight a persistent challenge in

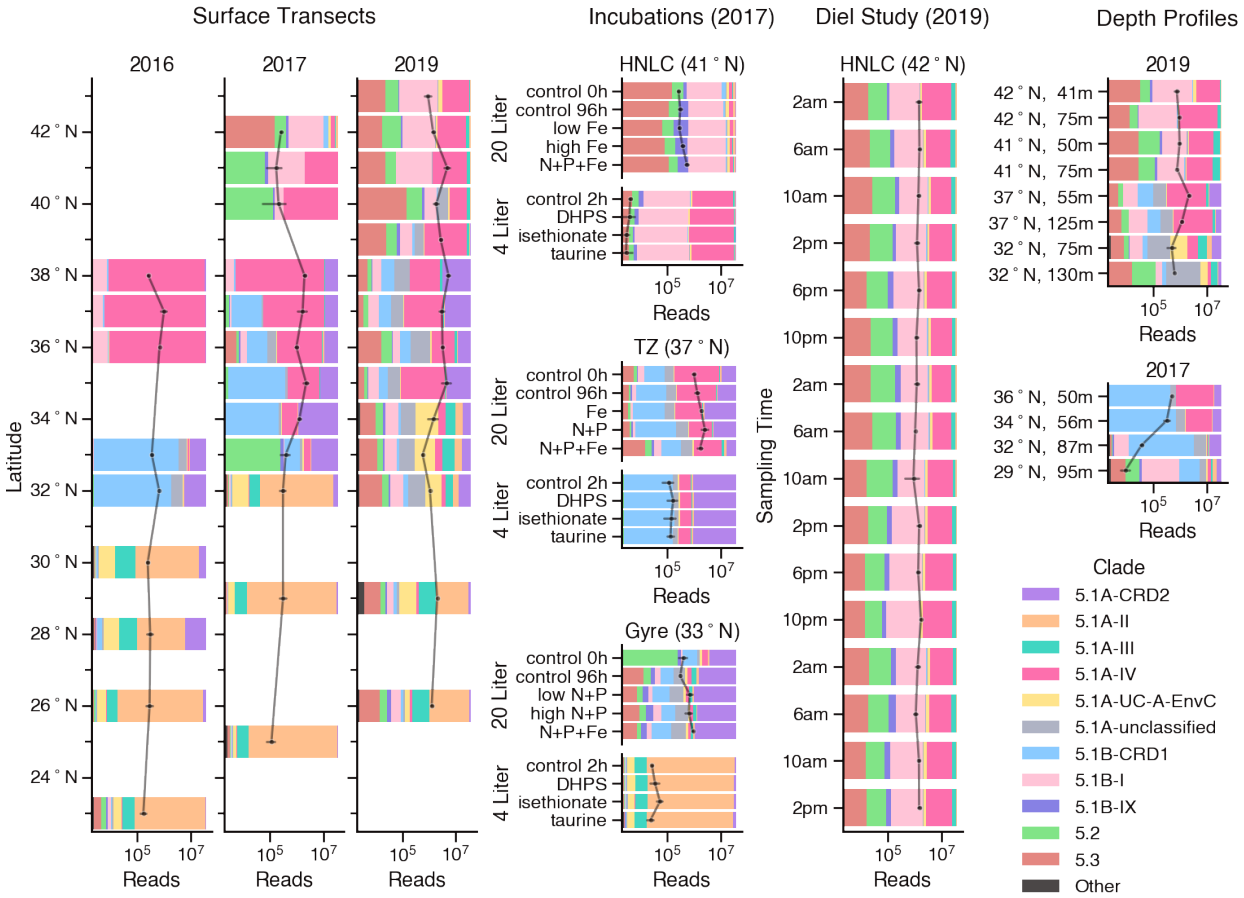

Figure 2.3: Composition of *Synechococcus* community transcript sequencing reads. Relative abundance of transcript reads mapped to *Synechococcus* clades. The size of each colored region is proportional to the fraction of sample reads mapped to the indicated clade (sample replicates averaged) with clade fractions summing to one in each sample, and samples grouped by dataset. X-axis corresponds to black line, showing aggregate count of *Synechococcus* transcript reads per sample with error bars indicating standard deviation of replicates.

the analysis of environmental metatranscriptomes: disentangling changes in gene expression from changes in organism abundance.

2.2.1 Normalization

To focus our analysis on patterns of differential gene expression and to adjust for differences between samples in community composition and sequencing depth, we normalized transcript abundance counts by clade. In both the *Prochlorococcus* and *Synechococcus* datasets, gene abundance profiles exhibited an overdispersed mean-variance relationship, indicative of a negative binomial distribution (Fig 2.4). We therefore normalized read counts using the variance stabilizing transform (vst) which employs a generalized linear model with negative binomial variance to stabilize overdispersion in count matrices [63, 64]. Normalization with vst produced residual transcript abundance values, which quantify the degree and direction to which each

transcript abundance diverged from expectation in each sample, given the distribution of abundance counts across all samples. These values can be described as transcription anomalies, which we interpret as instances of up-regulated or down-regulated gene expression. Genes that do not significantly vary between sampling conditions, after accounting for changes in organism abundance, will result in residual transcript abundance values near zero. Normalization successfully decoupled variance from mean, and the resulting values exhibited a similar range between clades [52]. We then arranged the normalized residual transcript abundance data into two tensorized datasets by aligning CyCOGs across clades. One tensor consisted of *Prochlorococcus* clades HLI, HLII, and LLI. The second tensor consisted of *Synechococcus* clades I, II, III, IV, CRD1, CRD2, and UC-A. *Prochlorococcus* clade LLVII and *Synechococcus* subclusters 5.2 and 5.3 were excluded from our analysis despite their apparent prevalence in some samples (Figs 2.2 and 2.3) because those transcript bins mapped to a low proportion of the clade pangenome, indicating a poor match between reference genomes and the sequenced metatranscriptomes.

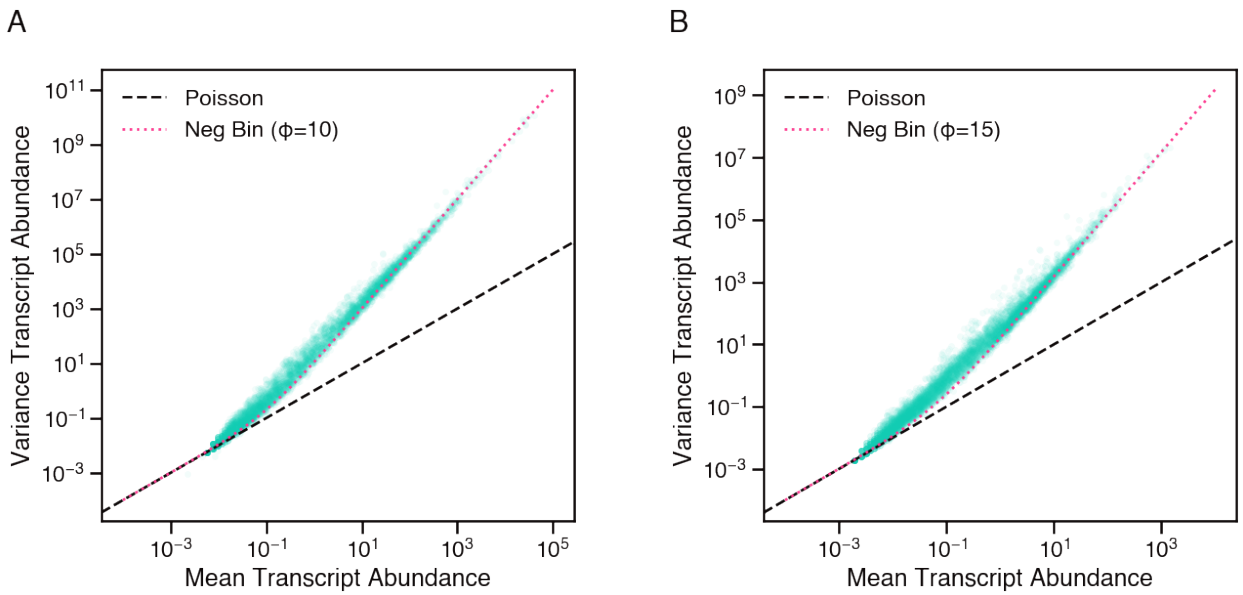

Figure 2.4: Raw transcript abundance counts exhibit an overdispersed mean-variance relationship. Variance (σ^2) vs. mean (μ) of (A) *Prochlorococcus* and (B) *Synechococcus* transcript abundance counts. Each point represents the transcript abundance profile of an individual gene (CyCOG) across samples mapped. As a visual guide, the mean-variance relationship of a Poisson distribution ($\sigma^2 = \mu$) is plotted as a dashed black line, and the mean-variance relationship of a negative binomial distribution ($\sigma^2 = \mu + \frac{\mu^2}{\phi}$) is plotted as dotted red line, parameterized with an appropriate overdispersion parameter, ϕ.

2.2.2 Model fitting to metatranscriptomic data

We applied Barnacle to the normalized *Prochlorococcus* and *Synechococcus* residual transcript abundance tensors to identify clusters of CyCOGs with correlated patterns of anomalous expression, attributable to

particular samples and clades. Best fit model parameters were identified using a cross-validated grid search strategy in combination with bootstrapping (see Methods). Bootstrapping allowed use of all samples for model fitting despite discrepancies in the number of replicates available for each sampling condition (see Methods and Fig 2.1). Among models fit without any applied sparsity penalty (λ=0.0), the minimum cross-validated SSE corresponded to an R of 15 components in both the *Prochlorococcus* and *Synechococcus* datasets. We selected the number of components based on the models with λ=0.0 as opposed to the overall minimum in SSE because among models fit with a sparsity coefficient of λ=10.0, SSE continually decreased with increasing R up to a value of $R = 1600$, a point at which computational power and practical considerations preclude efficient model analysis. A more granular search of λ values at the selected number of components ($R = 15$) led to best fit sparsity coefficients of $\lambda = 15.0$ for *Prochlorococcus* and $\lambda = 10.0$ for *Synechococcus*, based on the maximum cross-validated FMS. The *Prochlorococcus* model parameters ($R = 15, \lambda = 15.0$) resulted in a cross-validated SSE of 0.84 and a cross-validated FMS of 0.54, and the *Synechococcus* model parameters ($R = 15, \lambda = 10.0$) resulted in an SSE of 0.92 and an FMS of 0.53. The models fit with the selected parameters were then analyzed to evaluate the consistency and robustness of the resulting components.

2.2.3 Evaluation of component robustness

The FMS calculated during model fitting measures component consistency in the model as a whole. However, this metric does not measure the consistency and robustness of individual components. We therefore extracted and aligned components across 300 bootstrapped models to evaluate consistency, robustness to alternate parameters, and the percentage of variation attributable to each individual component (Fig 2.5). In each instance, the component-specific FMS scores were calculated based on comparison to a reference model parameterized with the identified best fit parameters (see Methods). Among the models with best fit parameters, a subset of components exhibited a tight distribution of component-specific FMS scores across bootstraps, indicating a high level of consistency between sample replicates (Fig 2.5, B and E). The majority of components exhibited a median FMS of 0.5 or greater, including 13 of the 15 *Prochlorococcus* components (Fig 2.5B), and 12 of the 15 *Synechococcus* components (Fig 2.5E). We also aligned components from models fit with suboptimal values of R and compared them against the best fit model components, to assess the robustness of components to alternate R parameters (Fig 2.5, A and D). Most components tended to exhibit a consistent median FMS score across the full range of R parameters in which they were detected. In general, components that accounted for a higher percentage of variation in the data (Fig 2.5, C and F) tended to be the "earliest" identified, detected in models with a small R parameter. Components that accounted for a lower percentage of variation tended to be picked up only by the more flexible models parameterized with a

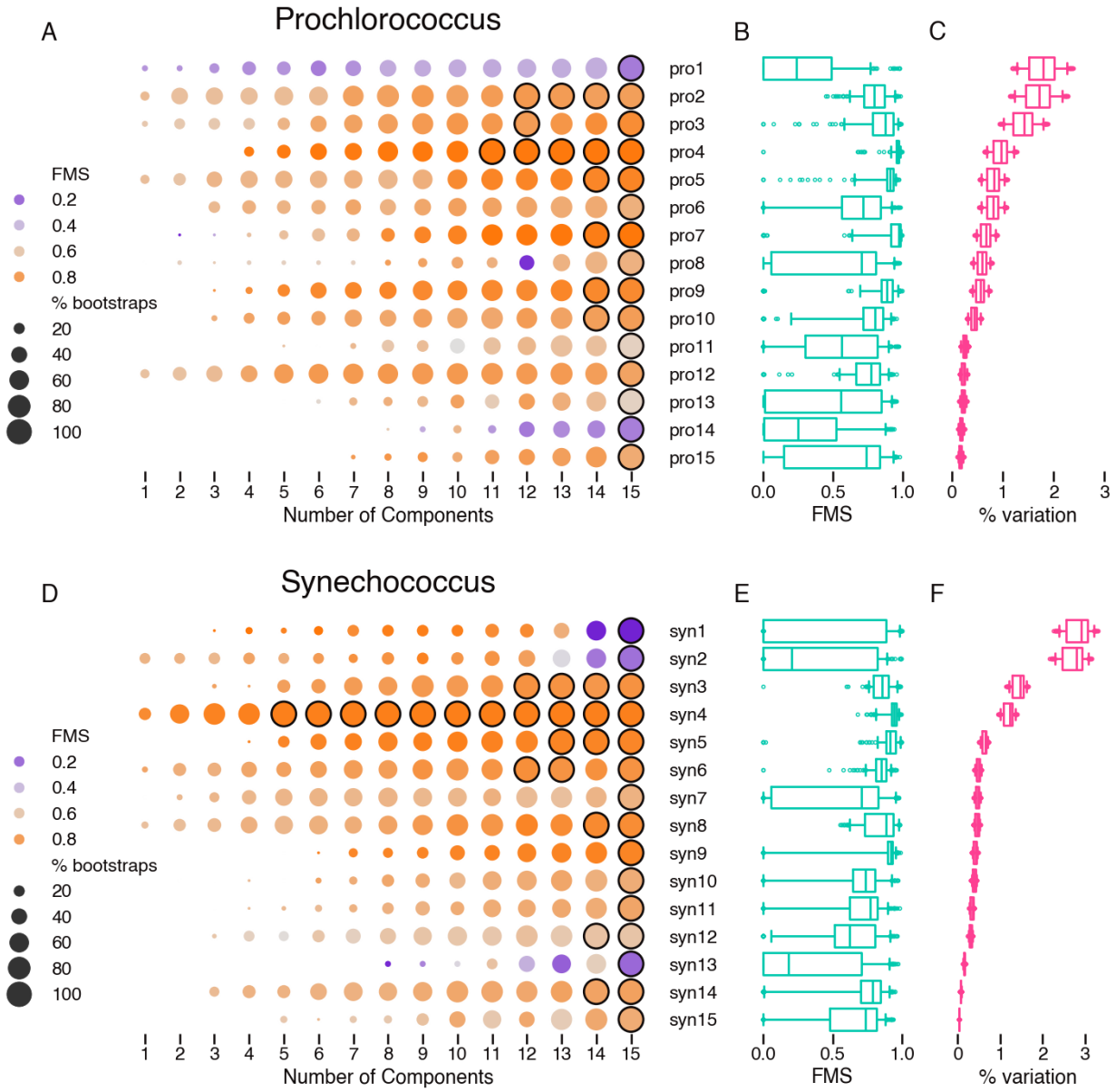

Figure 2.5: Individual components are robust to alternate values of R parameter. Analysis of components from sparse tensor decomposition models fit to (A-C) *Prochlorococcus* and (D-F) *Synechococcus* residual transcript abundance data. (A, D) Median factor match score (FMS) of components from models fit with different numbers of components (R) versus their corresponding best match component in the reference model with best fit parameters. Circle size represents the percentage of 30 bootstrap models found to contain the indicated component, circle color corresponds to median FMS, and circles outlined in black show components that were present in every bootstrap. (B, E) Distribution of component FMS across 300 bootstraps of best fit models. (C, F) Distribution across 300 bootstraps of percent variation represented by each component of best fit models. Boxes show inner 50th percentile centered on the median, whiskers indicate 5th and 95th percentiles, and outlying points are plotted individually.

greater number of components. Regardless of the minimum R parameter at which they were first detected, most components were subsequently detected at all values of R greater than this minimum, in an increasing percentage of bootstraps, and with a consistent FMS. This result indicated model stability, and suggested

that the patterns represented by these components are robust to model mis-specification.

We also evaluated the components for robustness to alternate parameterization of the sparsity coefficient. Components exhibiting a high median FMS with best fit parameters tended to maintain consistently high FMS scores across a wide range of alternate λ values [52]. Although the sizes of the gene clusters naturally differ across different values of λ, for most components a consistency of FMS scores across sparsity coefficients spanning almost two orders of magnitude indicated a stability in the signal modeled [52]. This reinforces the validity of our sparse modeling approach and suggests that the signal of many components is dominated by a core set of co-expressed genes that can be identified by applying an optimal level of sparsity. Based on these results we chose to focus the remainder of our analysis on the 13 *Prochlorococcus* components and 12 *Synechococcus* components that fell above a median FMS threshold of 0.5, removing components pro1, pro14, syn1, syn2, and syn13 from further analysis.

2.2.4 Interpretation of components

A cluster weight profile was generated for each of the remaining robust components by compiling the bootstrapped gene (CyCOG), taxon (clade), and sample weights. The profile delineates gene cluster membership, and summarizes the taxonomic and spatiotemporal patterns of cluster activity. At a 50% bootstrap threshold, the minimum *Prochlorococcus* cluster size was 4 CyCOGs, the maximum 100 CyCOGs, and the median cluster consisted of 22 unique CyCOGs. Of the 5,023 CyCOGs detected in the *Prochlorococcus* pangenome, 404 CyCOGs ($\sim 8\%$) were represented in at least one cluster. In *Synechococcus*, cluster size ranged from 16 to 107 CyCOGs with a median of 65 CyCOGs, altogether representing 638 out of 6,478 CyCOGs ($\sim 10\%$) detected in the *Synechococcus* pangenome. These statistics suggest that for both the *Prochlorococcus* and *Synechococcus* populations in this dataset, the most dramatic changes in expression are attributable to less than 10% of the expressed pangenome, perhaps corresponding to CyCOGs with a critical role in acclimating to specific environmental stresses encountered in our study.

The function of each CyCOG was inferred based on the majority annotation of member proteins, using several databases: the Kyoto Encyclopedia of Genes and Genomes (KEGG), Clusters of Orthologous Genes (COG), Pfam, and Tigerfam. Of the total 897 CyCOGs represented in at least one *Prochlorococcus* or *Synechococcus* cluster, we could assign a predicted function to 642 CyCOGS, while 255 CyCOGs had no assigned function. These CyCOG annotations, in conjunction with the KEGG pathway database, were used to analyze each cluster for over-represented biological functions (see Methods). All clusters were significantly enriched for at least one of 113 represented pathways, with the median cluster enriched for 12 pathways. These results signify biological coherence in the gene composition of each cluster, which we next explored in the context of the compiled taxon and sample weights.

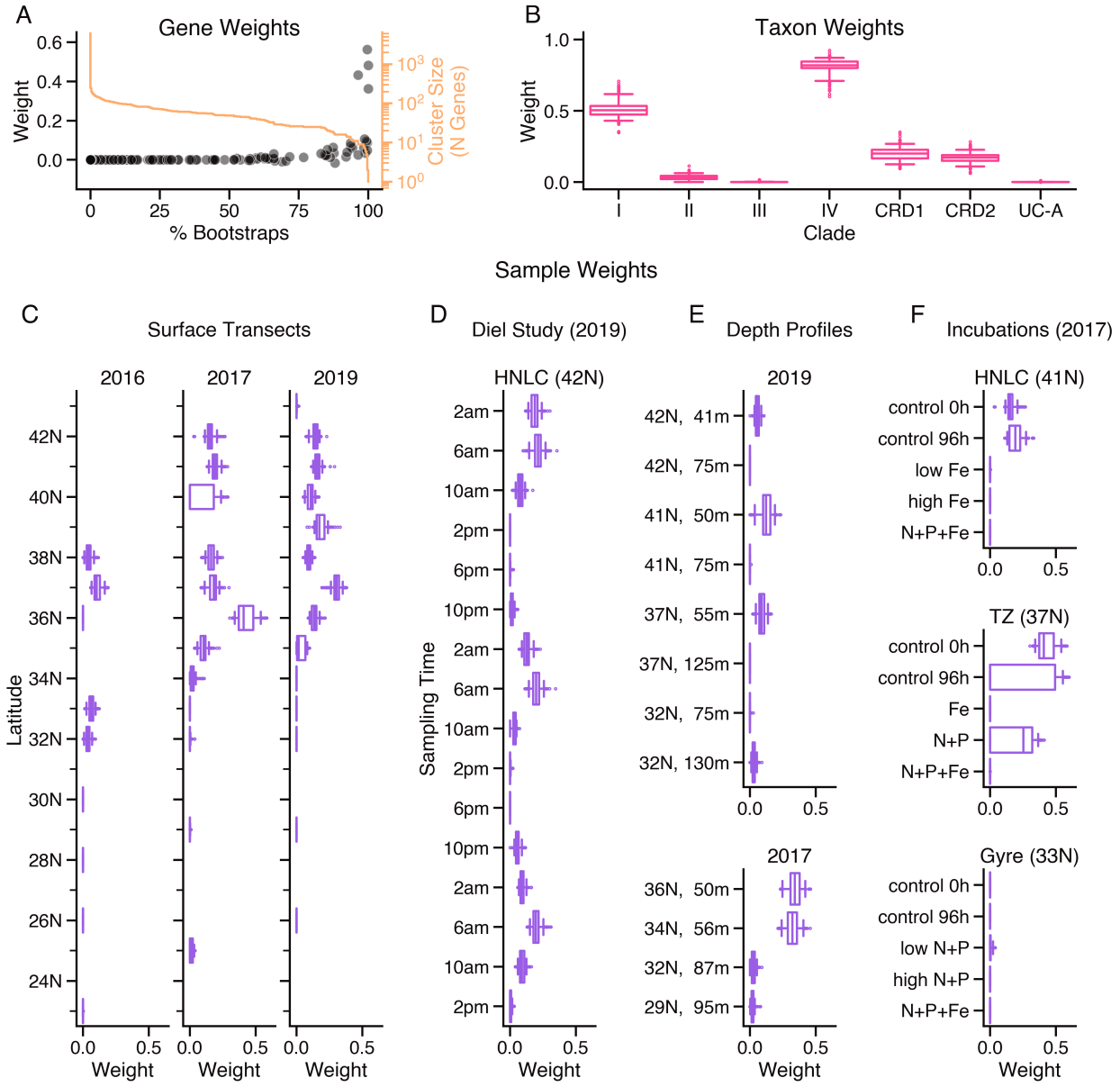

Figure 2.6: Example cluster profile of *Synechococcus* component syn8. (A) Median bootstrapped gene weight (black circles) of each CyCOG versus percentage of 300 bootstraps in which the CyCOG was included in component syn8 with a non-zero weight. Cluster size (number of CyCOGs) as a function of percent bootstrap threshold (orange line). Distribution of (B) taxon weights and (C-F) sample weights across 300 bootstraps. Boxes show inner 50th percentile centered on the median, whiskers indicate 5th and 95th percentiles, and outlying points are plotted individually. (C) Surface samples (≤15 meters) collected across 3 cruise years, binned by latitude. (D) Diel study samples binned by collection time of day. (E) Depth samples (>15 meters) collected in 2017 and 2019. (F) Incubation experiments in which indicated nutrients were amended to 20 liter seawater samples and transcriptomes collected after 96 hours. Component syn8 showed no significant activity in additional 4 liter incubations collected after 2 hours (data S2).

The cluster weight profiles, combined with CyCOG annotations and sample metadata, provide an intuitive interpretation of Barnacle components, as illustrated by *Synechococcus* component syn8 (Fig 2.6). Component syn8 encompassed 49 CyCOGs with nonzero median weights, and the four CyCOGs with the largest weights were detected in more than 95% of bootstraps (Fig 2.6A). Component syn8 also included six CyCOGs with negative median weights, demonstrating how Barnacle can incorporate within the same cluster genes with positively and negatively correlated expression profiles. The syn8 CyCOGs included genes involved in photosynthesis, carbon fixation, and adaptation to scarce iron. The taxon weights exhibited a tight distribution across bootstraps, indicating a consistent pattern in *Synechococcus* clades I and IV, and to a lesser degree clades CRD1, CRD2, and II (Fig 2.6B). The dataset incorporated samples collected under four different experimental protocols across the three cruises (Fig 2.1), and cluster syn8 exhibited anomalous expression in each sample group (Fig 2.6, C-F). In all three cruises the elevated expression of syn8 in surface waters was confined to the northern latitudes sampled, with little anomalous expression detected south of 32°N (Fig 2.6C). In 2019, a diel study at 42°N followed the same mass of water for 60 hours, sampling every 4 hours. Component syn8 exhibited a clear circadian pattern in this dataset, with a peak in expression each day at 6 am, and a nadir between 2 pm and 6 pm (Fig 2.6D). In depth profiles collected in 2017 and 2019, syn8 exhibited elevated expression at the latitudes where surface expression was also detected, and weights generally decreased with increasing depth (Fig 2.6E). In 2017, two sets of on-deck incubation experiments were conducted at three different latitudes. In the first set of experiments, 20 liter seawater samples were supplemented with nitrogen, phosphorous, and/or iron, and samples were collected for metatranscriptome sequencing after 96 hours of incubation. In the second set, 4 liter seawater samples were incubated with one of three sulfur-containing compounds and samples were collected for metatranscriptome sequencing after about two hours. Cluster syn8 expression was unremarkable in the 4 liter incubations and in the experiments conducted at 33°N (Fig 2.6F). However, in the incubation experiments conducted at 41°N, the syn8 cluster showed elevated expression in both the 0-hour and 96-hour control treatments, whereas no anomalous expression was detected in the three nutrient addition treatments, all of which included iron. Weights showed a wider spread across bootstraps in the transition zone experiments conducted at 37°N, indicating increased variability between these replicates. In this experiment the syn8 cluster showed elevated expression in the 0-hour and 96-hour controls, as well as the treatment amended with nitrogen and phosphorous, but no iron. In the two 96 hour treatments in which iron was added, syn8 showed no anomalous expression. In summary, the cluster profile of syn8 models a set of genes potentially related to photosynthesis, carbon fixation, and acclimation to low iron, common to *Synechococcus* clades I, IV, CRD1 and CRD2, with a joint pattern of anomalous expression in the northern part of the transect that exhibited a circadian peak at 6 am and was suppressed with the addition of iron.

2.2.4.1 Comparison between cluster weight profiles

To determine whether different components modeled similar processes, we compared cluster weight profiles between all pairs of *Prochlorococcus* and *Synechococcus* components (Fig 2.7). A majority of components corresponded to clusters of distinct, non-overlapping sets of genes (Fig 2.7A). Four groups of clusters (highlighted in color) had correlated gene weight profiles. Each of these clusters was specific to different clades (Fig 2.7B). For example, the gene weight profiles of clusters syn7, pro6, syn4, and syn10 represented similar sets of CyCOGs and yet their taxon weights differed: syn7 was primarily specific to *Synechococcus* clade CRD2, syn4 to clades I and IV, and syn10 to clades CRD1 and CRD2, whereas pro6 was specific to the HLI clade of *Prochlorococcus*. Several clusters had evident circadian expression patterns in the 2019 diel study samples, including syn8 (Fig 2.6), syn4, and syn6. Circadian expression was more difficult to discern for members of the other clusters, and we therefore inferred a daily peak expression time for each cluster by fitting a kernel density estimate to aggregated sampling times, weighted by the median component sample weights (see Methods). The first group of four clusters (syn7, pro6, syn4, syn10) exhibited a peak expression time around dusk, whereas the other six clusters (pro5, syn8, pro7, syn8, pro15, and syn12) had a peak expression time around dawn (Fig 2.7C).

Common physiological processes were apparent within each of the groups of clusters (Fig 2.7D). The clusters in the first group (syn7, pro6, syn4, and syn10) were each enriched in two KEGG pathways involved in aerobic respiration: oxidative phosphorylation and the pentose phosphate pathway. These clusters also had positive gene weights for all three subunits and several assembly proteins of cytochrome c oxidase, indicating elevated expression of the primary terminal oxidase. Two groups of clusters (pro5 and syn6, pro7 and syn8) were enriched in KEGG pathways describing carbon fixation in photosynthetic organisms, the pentose phosphate pathway, and glycolysis/gluconeogenesis. The KEGG pathway describing photosynthesis was also enriched in clusters pro5 and syn6. Photosynthesis was not enriched in pro7 and syn8, however syn8 showed increased expression of two photosystem II binding proteins, PsbQ and Psb28. In addition, high gene weights were assigned to the iron stress inducible genes *isiA* and *isiB*, both of which are understood to interact with photosynthesis processes [65, 66, 67, 68]. The fourth group consisted of clusters pro15 and syn12, which were both enriched for processes involved in nitrogen acquisition and assimilation, including ABC transporters in both clusters and nitrogen metabolism in syn12. Although cluster pro15 was not enriched for nitrogen metabolism, the four CyCOGs with nonzero median weights in the cluster were annotated as components of ammonium, urea, and taurine transporters. In summary, although the majority of clusters modeled distinct cellular processes, two independent analyses reiterated similar patterns of anomalous expression in gene modules related to respiration, photosynthesis, nitrogen acquisition, and acclimation to

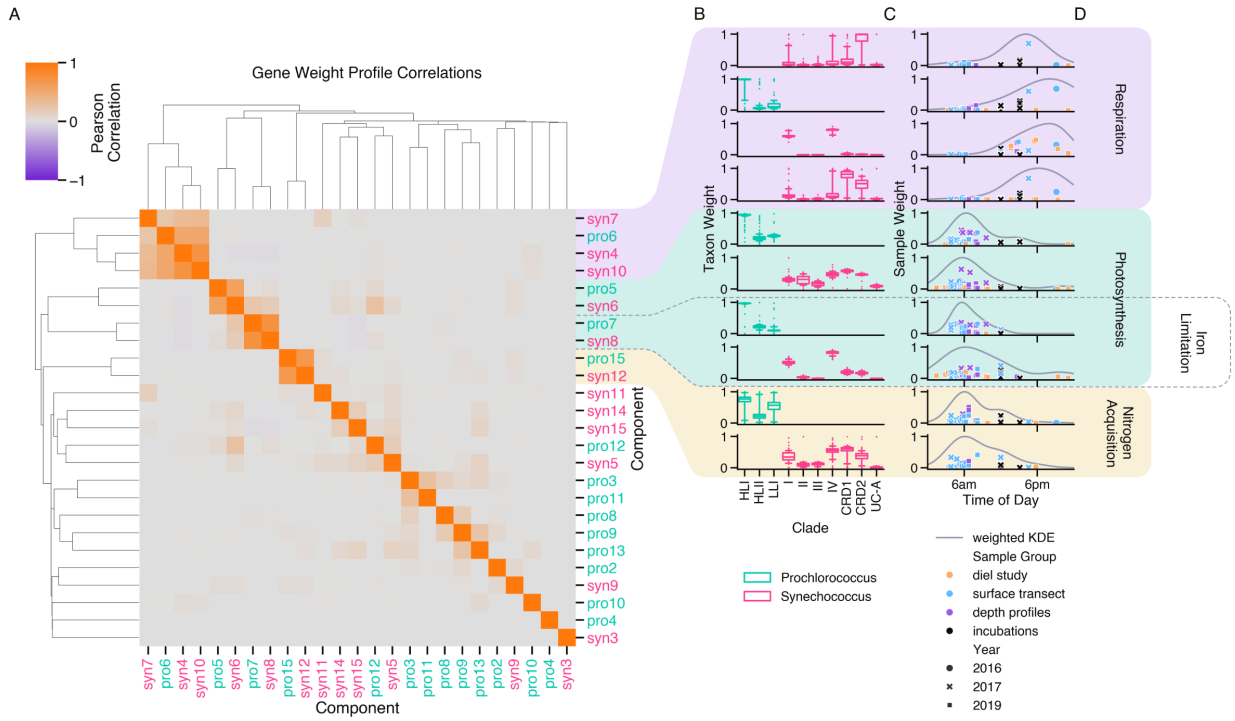

Figure 2.7: Subsets of *Prochlorococcus* and *Synechococcus* clusters exhibit similar gene content and temporal expression profiles. (A) Hierarchically clustered heatmap of pairwise Pearson correlations comparing gene weight profiles of *Prochlorococcus* and *Synechococcus* model components. Subsets of clusters with similar gene weight profiles are highlighted. (B) Distribution of bootstrapped taxon weight profiles of highlighted clusters. (C) Sample weights of highlighted profiles plotted by time of day of collection. In grey a weighted kernel density estimate indicates daily peak expression time. (D) Summary of predominant physiological processes of highlighted clusters as determined by enriched KEGG pathways and gene annotations.

scarce iron, indicating prominent, parallel shifts in these physiological activities across multiple co-occurring cyanobacterial populations in the North Pacific.

2.2.4.2 Comparison to laboratory gene expression clusters

To evaluate whether the co-expression patterns detected in field data recapitulated patterns seen in laboratory studies, we next compared the Barnacle-identified *Prochlorococcus* and *Synechococcus* clusters against previously identified clusters from a diel study of *Prochlorococcus* strain MED4, grown under nutrient replete culture conditions (Zinser et al., 2009) [69]. We mapped the MED4 genes to their corresponding CyCOG and calculated a weighted F1 score that compared the gene content of each pair of Zinser- and Barnacle-identified clusters in an all-by-all manner. After correcting for multiple comparisons, 4 of the 18 Zinser clusters overlapped significantly with 9 of the 25 Barnacle clusters (Fig 2.8). The four respiration clusters pro6, syn4, syn7, and syn10 (Fig 2.7) were significantly similar to Zinser cluster 5, which Zinser et al. found

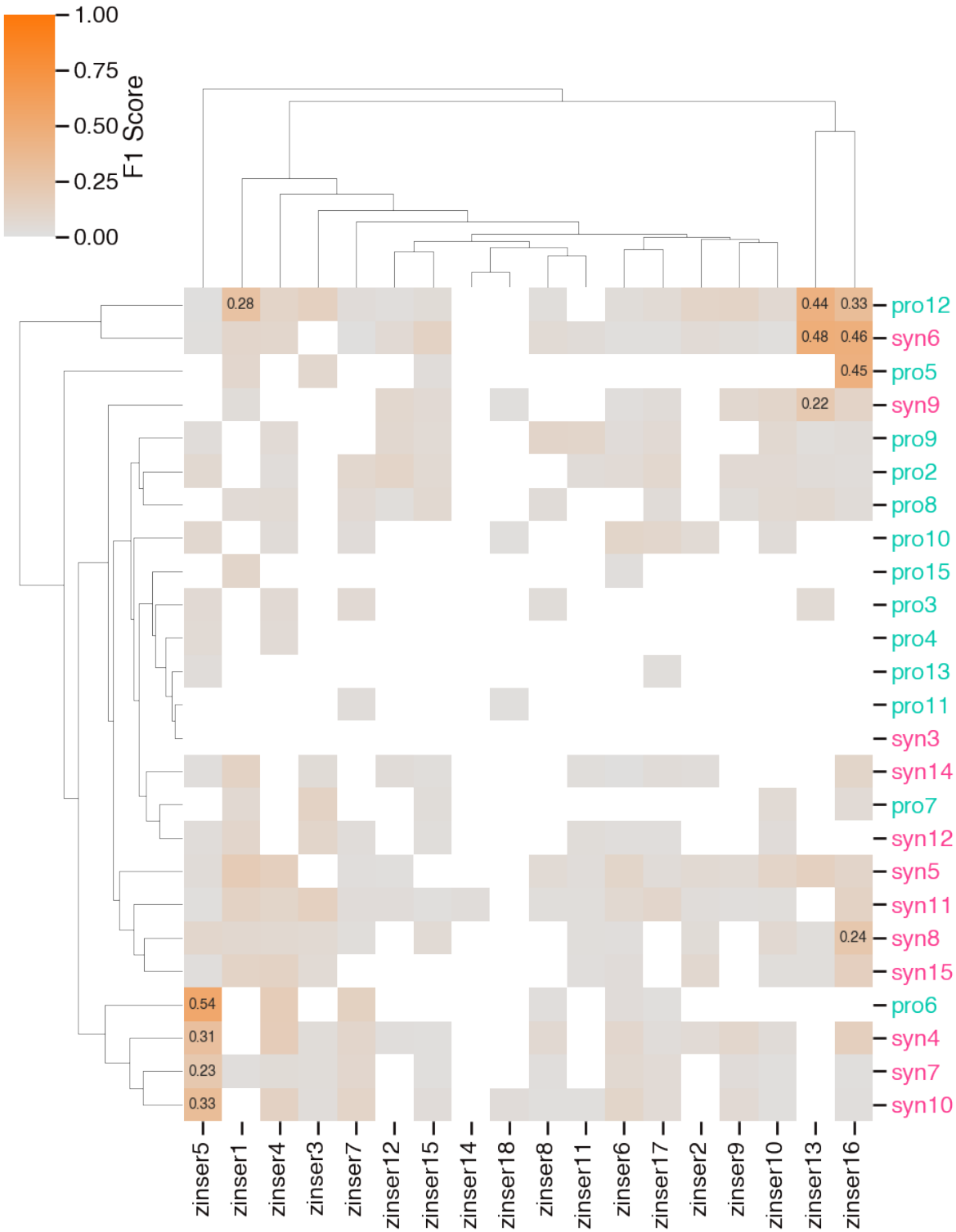

Figure 2.8: Comparison of Barnacle cluster gene membership to previously reported diel clusters of cultured *Prochlorococcus* MED4. Hierarchically clustered heatmap of F1 scores comparing CyCOG composition of Barnacle clusters against clusters previously derived from transcriptomes of *Prochlorococcus* MED4 grown in culture over a simulated day/night cycle [69]. Clusters without any overlapping CyCOGs are left blank and significantly similar clusters are annotated with the F1 score (adjusted $p < 0.05$).

to be enriched for respiratory terminal oxidases and degradation of proteins, peptides, and glycopeptides. The peak expression time of the Barnacle respiration clusters occurred around 6 pm (Fig 2.7C) and the Zinser cluster displayed a peak expression time of 5:30 pm ± 0.4 hours, in both instances around sunset. Barnacle photosynthesis clusters pro5, syn6, and syn8, along with cluster pro12 displayed a peak expression time around 6 am (Fig 2.7) and were significantly similar to Zinser cluster 16 (Fig 2.8), which had a peak expression time of 5:30 am ± 0.5 hours and was enriched for ATP synthase and CO_2 metabolism. Barnacle clusters pro12, syn6, and syn9 significantly overlapped with Zinser cluster 13 (Fig 2.8), a cluster enriched for ribosomal proteins with a peak expression time of 3:24 am ± 0.4 hours. The most significantly enriched KEGG pathway in these three Barnacle clusters corresponded to ribosomal proteins (data S3) and the peak expression time was approximately 6 am. The second most significantly enriched KEGG pathway in pro12 corresponded to photosynthesis. The pro12 cluster also overlapped significantly with Zinser cluster 1 (Fig 2.8), a cluster enriched for photosystems I and II and with a peak expression time of 8:18 am ± 0.7 hours. The remarkable similarity in gene content and expression peaks between the Barnacle clusters identified from environmental data and the clusters identified in laboratory-grown MED4 reiterates the power of both our modeling approach and laboratory studies with model organisms.

2.2.5 Acclimation to nutrient scarcity in the North Pacific

The North Pacific Ocean is characterized by iron limitation of photosynthetic organisms in subarctic regions and nitrogen limitation in the subtropical regions [60]. Our comparison of cluster gene weight profiles revealed one pair of clusters related to nitrogen acquisition (pro15, syn12) and a second pair of clusters (pro7, syn8) containing genes linked to iron stress (Fig 2.7). Within each pair, the two clusters shared similar sample weight profiles, with pro7 and syn8 showing larger weights in the northern section of the transect and pro15 and syn12 showing larger weights in the southern section. These observations led us to deduce that the four clusters model processes relevant to acclimation and ecological partitioning along nutrient gradients in the North Pacific, providing an opportunity to study the in situ physiological effects of these environmental stresses by examining cluster gene content and sample weights in more detail (Fig 2.9).

2.2.5.1 Nitrogen acclimation response

Prochlorococcus and *Synechococcus* populations in the North Pacific uptake ammonium and urea as their primary nitrogen source [78], and subpopulations of both genera have been shown to assimilate nitrate [78, 79]. A close examination of the CyCOGs in clusters pro15 and syn12, and their associated gene weights (all of which were positive), revealed elevated expression of four genes common to both clusters: *amt1*, *tauA*, *urtA*, and *urtB* (Fig 2.9B). The *urtA* and *urtB* genes encode two components of a urea ABC-transporter

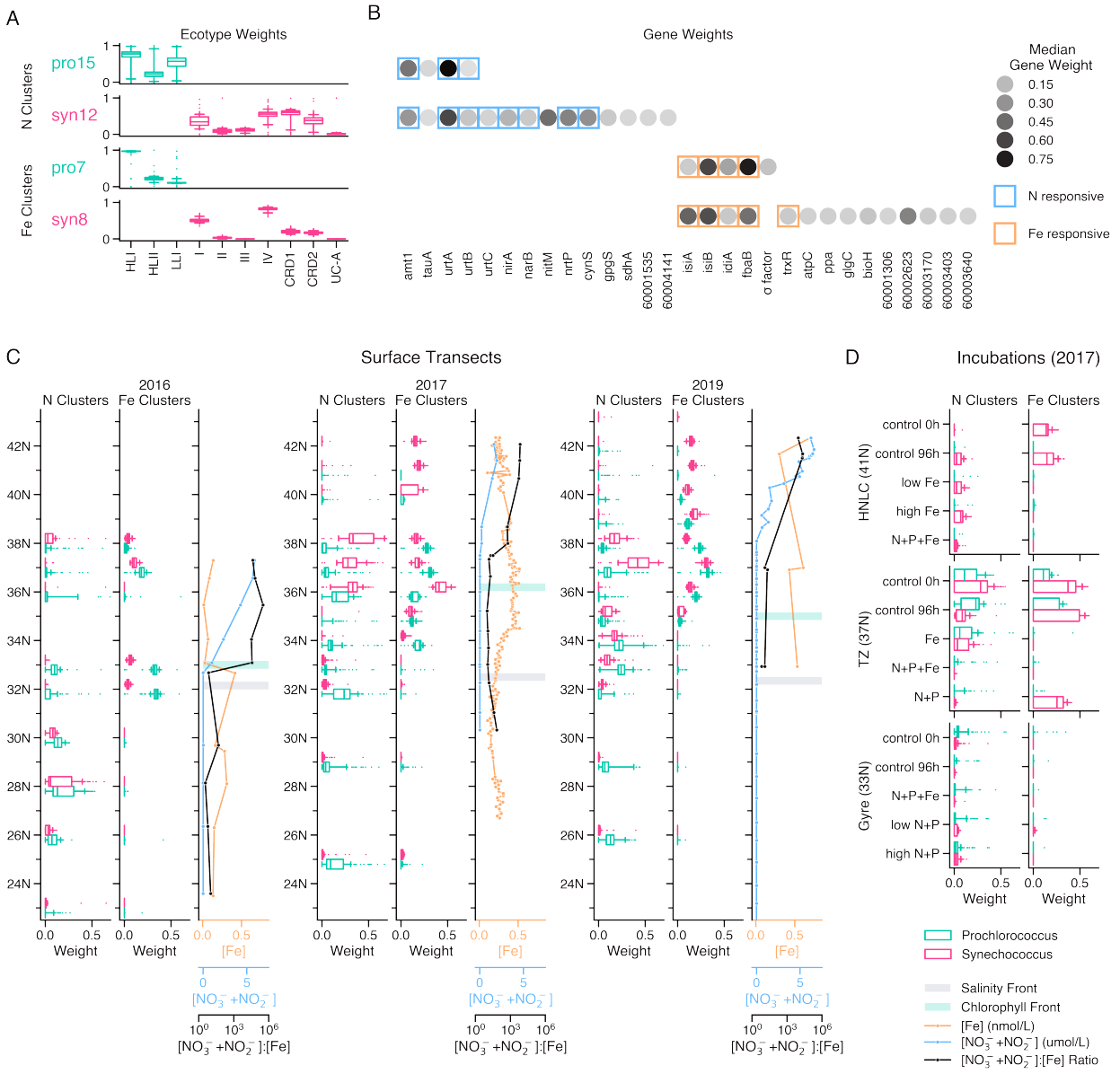

Figure 2.9: Independent *Prochlorococcus* and *Synechococcus* clusters signal acclimation to scarce nitrogen and iron overlapping in the subtropical-subarctic transition zone. Bootstrapped component weight profiles of *Prochlorococcus* (green) and *Synechococcus* (pink) clusters enriched for genes related to nitrogen scarcity (pro15, syn12) and iron scarcity (pro7, syn8). (A) Distributions of cluster taxon weights. (B) Top cluster genes with absolute median weight ≥ 0.04. Boxes indicate genes previously reported to have significantly altered expression in response to nitrogen deprivation (blue) [70, 71, 72] and iron deprivation (orange) [65, 73, 74, 75, 76]. (C) Surface sample weight distributions for three years of cruises, binned by latitude, with nitrogen-responsive cluster weights in leftmost columns and iron-responsive cluster weights in middle columns. Rightmost columns show the nitrogen-to-iron ratio (black) calculated using measurements of surface dissolved iron concentration (orange) and dissolved nitrate plus nitrite (blue). Salinity and chlorophyll fronts as reported by Juranek et al. (2020) are indicated by grey and green bars, respectively [77]. (D) Distributions of sample weights from 20 liter nutrient amendment experiments measured after 96 hours of incubation and conducted at three latitudes in 2017. Boxes show inner 50th percentile centered on the median, whiskers delineate 5th and 95th percentiles, and outlying points are plotted individually.

system and *amt1* encodes an ammonium transporter. Increased expression of *amt1* and *urtABCDE* has been observed in response to nitrogen deprivation in marine *Prochlorococcus* [71] and *Synechococcus* [72]. The *tauA* gene encodes the substrate-binding component of an ABC-transporter system that has recently been proposed to import guanidine in marine cyanobacteria from low nitrogen environments [80]. The *Prochlorococcus* cluster pro15 was limited to these four CyCOGs, whereas the *Synechococcus* cluster syn12 consisted of a total of 20 CyCOGs. The additional syn12 CyCOGs included nitrite reductase (*nirA*), nitrate reductase (*narB*), the formate/nitrite transporter (*nitM*) and the bi-specific nitrate/nitrite permease (*nrtP*), all of which are involved in nitrate and nitrite assimilation (Fig 2.9B). In *Synechococcus* the expression of *nirA*, *narB*, and *nrtP* is controlled by the universal nitrogen response regulator NtcA, and up-regulated in response to nitrogen deprivation [70]. Cyanate lyase (*cynS*), included in syn12, is also expressed in response to ammonium deprivation in an NtcA-dependent manner [81]. A key enzyme in the synthesis of glucosylglycerate, *gpgS*, was also included in syn12. Glucosylglycerate is a compatible solute that has been found to accumulate in *Synechococcus* under nitrogen limitation, possibly replacing glutamate as a negatively charged counterion [82]. We detected additional genes related to nitrogen metabolism with bootstrap values that fell below the 50% threshold, including CyCOGs corresponding to NtcA (CyCOG 60000127) and glutamine synthase (CyCOG 60000563). A small fraction of pro15 bootstraps also included nitrate and nitrite reductases, possibly reflective of the patchy distribution of *nirA* and *narB* among *Prochlorococcus* strains at the sub-clade level [79].

2.2.5.2 Iron acclimation response

A deeper evaluation of the gene content of clusters pro7 and syn8 revealed four shared CyCOGs with large positive weights, all of which are implicated in acclimation to iron scarcity in cyanobacteria (Fig 2.9B). Three of the shared CyCOGs are canonical markers of iron stress: *isiA* encodes a chlorophyll binding protein that forms an antenna structure around photosystem I in iron-limited cyanobacteria [65, 67, 68], *isiB* encodes flavodoxin, an iron-free analog of the iron-containing electron transfer protein ferredoxin [65, 66], and *idiA* is hypothesized to encode either a protein that protects photosystem II against oxidative stress during iron limitation [83] or a component of an ABC-type Fe(III) transporter [84]. All three genes have been shown to be up-regulated in response to iron limitation in both *Prochlorococcus* [73, 74] and *Synechococcus* [65, 75, 76]. The fourth shared gene, *fbaB*, encodes a class I fructose bisphosphate aldolase (FBA), an enzyme involved in several carbon metabolism pathways including carbon fixation. There are two variants of FBA: class II FBAs require a divalent iron or zinc cation as a cofactor, whereas class I FBAs do not require a metal cofactor [85] and have been found to be up-regulated in response to iron limitation in *Synechococcus* [76] and diatoms [86]. The class I FBA (*fbaB*; CyCOG 60001290) was assigned positive weights in both pro7

and syn8, indicating elevated expression, whereas the class II FBA (CyCOG 60001287) was not included in either cluster.

As with the nitrogen-related clusters, the iron-related *Synechococcus* cluster (syn8) was larger than the iron-related *Prochlorococcus* cluster, encompassing a total of 49 CyCOGs, of which 17 are uncharacterized. The annotated CyCOGs specific to *Synechococcus* that showed the highest degree of elevated expression (Fig 2.9B) fell into one of four functional categories: oxidative stress response (*trxR*), ATP production (*atpC*, *ppa*), glycogen production (*glgC*), and biotin biosynthesis (*bioH*). The gene product of *trxR* is thioredoxin reductase, an enzyme involved in mitigating oxidative stress that has shown increased expression in *Synechococcus* subjected to low iron conditions [76]. To a lesser degree, the antioxident peroxiredoxin (*ahpC*) also showed elevated transcription in syn8; elevated levels of peroxiredoxin have been observed in the proteome of iron-limited cultures of clade II *Synechococcus* [87]. Similarly, in addition to *glgC*, the genes *glgA* and *glgB* were included in syn8 with small positive weights, meaning the complete glycogen synthesis pathway exhibited increased expression in syn8. Increased glycogen production has been observed in laboratory studies of *Synechococcus* facing iron scarcity [88], and may be connected to stress response in cyanobacteria [89]. Finally, biotin biosynthesis requires the precursor pimelate thioester, which can be synthesized through either an iron-dependent (*bioI-bioW*) or an iron-independent (*bioC-bioH*) pathway [90]. The elevated expression of *bioH* in cluster syn8 is consistent with a pattern of swapping out iron containing proteins with iron-free analogs during iron limitation, and may reflect another instance of this strategy to reduce the cellular iron quota of *Synechococcus*.

Both the pro7 and syn8 clusters included additional CyCOGs that exhibited a smaller degree of differential expression and yet were consistently represented in bootstraps, thus providing additional support for trends inferred from the CyCOGs with larger gene weights. Two additional CyCOGs potentially involved in iron acquisition had positive gene weights: CyCOG 60002762 in syn8 was annotated as a hemoglobin/transferrin/lactoferrin receptor protein (KEGG Orthology K16087), and *gap2* in pro7 encodes glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Although GAPDH is primarily involved in the Calvin cycle, it has also been shown to moonlight as a siderophore receptor in diverse bacteria [91]. In addition to *gap2*, a number of other Calvin cycle genes showed elevated expression in both clusters: *rpe* in pro7, and in syn8, *rbcS*, *csoSCA*, *csoS2*, *prkB*, and *glpX*. In syn8 the increased expression of carbon fixation genes was accompanied by elevated expression of photosystem II binding proteins (PsbQ, Psb28) and genes involved in phycobilin biosynthesis (*hemE*, *hemH*, *hmuO*), indicating that the *Synechococcus* iron scarcity response includes an adjustment of several structures and pathways related to photosynthesis. The syn8 cluster also included genes that displayed negative gene weights: three cytochrome c oxidase genes, indicating decreased expression of the respiratory terminal oxidase, and the chlorophyll biosynthesis gene *clhN*. Cyanobacteria

encode two protochlorophyllide reductases; *clhN* encodes the iron-sulfur subunit of one while its analog has no iron requirement [92]. Finally, in line with an elevated oxidative stress response, syn8 included increased expression of an under-characterized CyCOG (60003403) that shares homology with DoxX, a component of a membrane-bound protein complex that co-localizes a thiol oxidoreductase with a superoxide dismutase in *M. tuberculosis* [93]. Altogether the gene weights of clusters pro7 and syn8 suggest a broad program of transcriptional modulation to import more iron, decrease cellular iron quota, adjust photosynthesis, and mitigate oxidative stress.

2.2.5.3 Latitudinal trends in nutrient acclimation

Surface sample weights indicated that the transcriptional patterns modeled by clusters pro7, syn8, pro15, and syn12 were geographically bounded, consistent across study years, and echoed in both *Prochlorococcus* and *Synechococcus* populations (Fig 2.9C). The nitrogen acquisition clusters pro15 and syn12 showed consistent low levels of elevated expression in samples collected in the subtropical gyre and further north, to about 38°N (Fig 2.9C). In contrast, the iron acclimation clusters pro7 and syn8 exhibited elevated expression north of the subtropical gyre, with syn8 expression extending to around 42°N, the northernmost stations sampled on the 2017 and 2019 cruises (Fig 2.9C). The *Prochlorococcus* iron acclimation cluster pro7 was not detected in the northernmost surface samples, likely because this cluster primarily represents the HLI clade (Fig 2.9A), which did not make up a significant portion of the *Prochlorococcus* community in those samples (Fig 2.2). Rather, at these latitudes the *Prochlorococcus* community was dominated by the LLI and LLVII clades, the former noteworthy for a high frequency of genes involved in iron acquisition via siderophore uptake [94]. In the transition zone (TZ) between the subtropical and subarctic gyres, expression of both the nitrogen and iron acclimation clusters was elevated in both *Prochlorococcus* and *Synechococcus* populations (Fig 2.9C). The ratio of dissolved nitrate to dissolved iron increased north of 33°N in 2016 and north of 37°N in 2017 and 2019. In all three years this inflection point coincided with the region in which the iron and nitrogen acclimation clusters showed simultaneous elevated expression in both the *Prochlorococcus* and *Synechococcus* populations.

To investigate whether the in situ gene expression patterns observed were consistent with experimental evidence of nutrient limitation, we examined sample weights associated with the 2017 nutrient amendment incubation experiments (Fig 2.9D). At the NPSG station (33°N), the nitrogen clusters (pro15, syn12) had non-zero sample weights, indicating a low level of elevated expression, whereas the weights for the iron acquisition clusters (pro7, syn8) were zero. Conversely, at 41°N the *Synechococcus* iron cluster syn8 showed elevated expression in both the 0-hour and 96-hour controls, whereas the *Synechococcus* nitrogen cluster syn12 was associated with a zero weight in the 0-hour control sample and slightly elevated expression in the 96-hour

incubations, although less so in the treatment amended with supplemental nitrogen. The *Prochlorococcus* clusters showed no activity in this set of samples since *Prochlorococcus* clade HLI was not abundant at this latitude (Fig 2.2). In the transition zone incubations conducted at 37°N, all four clusters showed elevated expression in the 0-hour and 96-hour controls (Fig 2.9D). Moreover, at this site treatments amended with nitrogen resulted in sample weights of zero for the nitrogen clusters, and treatments amended with iron resulted in sample weights of zero for the iron clusters. These results imply that expression of clusters pro7 and syn8 was specifically suppressed with the addition of iron, while pro15 and syn12 expression was suppressed with the addition of nitrogen. Furthermore the results provide evidence of a North Pacific gradient of nitrogen limitation in the subtropical gyre and iron limitation in the subarctic gyre, and indicate a region of simultaneous nutrient stresses experienced by the *Prochlorococcus* and *Synechococcus* populations at the transition zone boundary between these two ecosystems.

2.3 Discussion

The patterns uncovered by applying Barnacle to the compiled metatranscriptomic dataset underscore the preeminence of nitrogen and iron scarcity in structuring North Pacific microbial ecosystems, and suggest novel mechanisms by which *Prochlorococcus* and *Synechococcus* populations adapt to these prevalent stressors. Two of the most robust components identified in our analysis, pro7 and syn8, suggest that the acclimation response to scarce iron exhibits a similar pattern in *Prochlorococcus* and *Synechococcus*, with important physiological differences distinguishing the two lineages. As an efficient electron transfer agent, iron is essential to the functioning of many core cellular processes, including photosynthesis and respiration. Consequently, iron scarcity poses a serious obstacle to photosynthetic growth and is one of the main limiting nutrients across much of the surface ocean [60]. Overcoming iron scarcity requires coordinated adjustment of intertwined cellular processes. Three physiological categories of cyanobacterial acclimation to iron limitation were proposed by Straus (1994) [95]: 1) acquisition of more iron, 2) compensation via the replacement of iron-requiring proteins with iron-free analogs, and 3) retrenchment, which entails a reduction of cell structures, particularly photosystem I and other iron-rich components of photosynthetic electron transport (PET) [96]. Additionally, iron limitation poses an increased risk of oxidative damage, in part because retrenchment and compensation decrease the number of iron centers in PET, which makes saturation of the PET chain more likely and increases the likelihood that electrons will escape and form reactive oxygen species [97]. As such, a fourth category of cyanobacterial acclimation to low iron was later added by Michel and Pistorius (2004) to distinguish the role that some iron deficiency induced genes play in mitigating oxidative stress [83]. All four categories of cyanobacterial response to scarce iron are reflected in the physiological

states modeled by clusters pro7 and syn8. As such, our results recapitulate well-established patterns of iron stress response, and additionally reveal new potential response mechanisms and contrasts in iron physiology between *Prochlorococcus* and *Synechococcus*.

Prochlorococcus cluster pro7 encompassed seven CyCOGs with elevated expression (Fig 2.9B), each with a putative role in iron acquisition, compensation, retrenchment, or protection from oxidative stress. Three CyCOGs correspond to well-established gene markers of iron deficiency: *isiA*, *isiB* (flavodoxin), and *idiA*. The *isiA* and *idiA* gene products are hypothesized to protect photosystems I and II from oxidative damage, respectively [83]. The *idiA* gene product may also be involved in iron acquisition as it displays homology to the periplasmic component of the bacterial ferric uptake transporter [83]. Flavodoxin expression is a canonical example of compensation, functionally replacing the iron-sulfur electron shuttle ferrodoxin with an iron-free analog [95, 73, 74, 75]. The elevated expression of an iron-free fructose bisphosphate aldolase (FBA) analog in pro7 may be another example of this compensation pattern. FBAs are involved in several carbon metabolism pathways, including the Calvin cycle, as are two of the remaining CyCOGs in pro7: *rpe* encoding ribulose-5-phosphate 3-epimerase (RPE) and *gap2* encoding a glyceraldehyde-3-phosphate dehydrogenase (GAPDH). Our analysis accordingly found that carbon fixation was the most significantly enriched KEGG pathway in pro7. The rate of carbon fixation limits the rate of photosynthesis in cyanobacteria [98] and so increasing the expression of critical Calvin cycle enzymes could serve to compensate for low-iron-induced retrenchment by maximizing photosynthetic efficiency. Maximizing the carbon fixation rate could also serve to reduce the formation of reactive oxygen species by maximizing the rate at which electrons flow out of PET and into the Calvin cycle, thus relieving saturation of the PET chain. Additionally, RPE employs an iron atom in its active site, and both it and GAPDH are inactivated by oxidation [99, 100]. These enzymes may be points of the Calvin cycle vulnerable to oxidative damage, and thus their up-regulation may serve as a protective response to the increased oxidative stress that accompanies iron limitation. Finally, pro7 includes one CyCOG corresponding to a sigma factor. The presence of a single sigma factor within this cluster suggests it may be involved in coordinating the *Prochlorococcus* transcriptional program across these four categories of acclimation to low iron.

Synechococcus cluster syn8 exhibits many of the same hallmarks of low iron acclimation seen in the *Prochlorococcus* cluster. Iron deficiency marker genes *isiA*, *isiB*, and *idiA* all show a similar degree of over-expression, as does *fbaB*, encoding the iron-free class I FBA enzyme (Fig 2.9B). However, our results indicate that *Synechococcus* acclimation to low iron is more involved than that of *Prochlorococcus*: syn8 encompasses a total of 49 CyCOGs as compared to the 7 CyCOGs of pro7. This difference may reflect the evolutionary streamlining of *Prochlorococcus* metabolism in comparison to the greater metabolic flexibility of *Synechococcus* [101], which allows it to thrive across a greater range of environmental conditions [55].

Prochlorococcus metabolism is likely already well-equipped to cope with nutrient scarcity, and so adapting to low iron conditions requires relatively minor tuning of its transcriptional program. In contrast, while *Synechococcus* possesses the genetic potential to thrive across a range of nutrient concentrations, actualizing acclimation to low iron entails broader metabolic restructuring. Here the inferential power of Barnacle components is on full display, allowing us to detect the cohort of genes up-regulated in concert with markers of low iron stress, and to extrapolate from their collective activity the mechanism physiological acclimation in *Synechococcus*.

The shifts in gene expression specific to cluster syn8 appear primarily geared towards maximizing the efficiency of photosynthesis and downstream metabolic processes, and mitigating oxidative stress. Two CyCOGS encoding enzymes involved in oxidative stress response showed elevated expression: thioredoxin reductase (*trxR*; Fig 2.9B) and peroxiredoxin (*ahpC*), both of which have been previously observed up-regulated in iron-limited *Synechococcus* [76, 87]. Seven up-regulated CyCOGs are involved in carbon fixation, and another two are involved in ATP synthesis. Similar to RPE in *Prochlorococcus*, many of these enzymes are susceptible to oxidative damage, and their increased expression may serve as a "spare parts" repository, speeding up the repair of damaged components and reducing potential metabolic bottlenecks resulting from slow repair. Similarly, genes encoding the photosystem binding proteins PsbQ and Psb28, up-regulated in cluster syn8, are associated with photosystem II repair in cyanobacteria, following oxidative inactivation [102, 103]. Elevated Psb28 abundance has also been observed in the proteome of clade II *Synechococcus* subjected to iron limitation [87]. Increased expression of genes encoding the enzymes uroporphyrinogen decarboxylase (*hemE*), ferrochelatase (*hemH*), and heme oxygenase (*hmuO*) suggests an increased investment in the construction of phycobilisomes, as each enzyme catalyzes a key branch-point reaction in phycobilin biosynthesis. In addition to harvesting light for photosynthesis, phycobilisomes mediate non-photochemical quenching in *Synechococcus*, dissipating excess light energy via state transitions [104] and interaction with the orange carotenoid protein [105]. The investment in phycobilisome synthesis may enhance a cell's capacity to divert light energy from PET and reduce the potential formation of reactive oxygen species. At the opposite end of electron transport, syn8 gene weights indicated increased expression of genes required for the glycogen biosynthesis pathway (*glgABC*), and decreased expression of the respiratory terminal oxidase. Glycogen metabolism is implicated in boosting the efficiency of photosynthesis and carbon fixation in *Synechococcus*, particularly during the transition from dark to light [89]. Glycogen is also a terminal sink for photosynthetic electrons, and increasing its production may help relieve pressure on the PET chain. The decrease in respiration is best interpreted in the context of the diel pattern of the syn8 expression profile, which peaked around dawn (Fig 2.6). Our data (Fig 2.7) and previous studies [69] indicate that cyanobacterial respiration peaks at dusk; the observed down-regulation of respiration genes in syn8 may indicate a

deepening of the diel segregation of photosynthesis and respiration in response to scarce iron, perhaps to minimize additional electron inputs to the components of the electron transport chain that are shared by photosynthesis and respiration [106]. Altogether the syn8 cluster touches multiple stages of photosynthesis, from light harvesting to ATP synthesis, carbon fixation, and glycogen production, suggesting a program of acclimation to low iron in *Synechococcus* that maximizes photosynthetic electron flux, enhances diversion routes for excess reducing power, and mitigates oxidative stress. Intriguingly, the transcriptional responses in *Synechococcus* uncovered here are reminiscent of streamlining events proposed to underlie the evolutionary adaptation of *Prochlorococcus* lineages to low nutrient conditions [101].

At the ecosystem scale, the sample weights associated with iron acclimation clusters pro7 and syn8 showed a latitudinal trend of elevated expression in the north and lower expression in the south, inverse to the sample weights of nitrogen assimilation clusters pro15 and syn12 (Fig 2.9, C and D). Just as pro7 and syn8 genes *isiA* and *isiB* have been used as indicators of low iron stress, the genes *urtABC* (urea transporter), *nirA* (nitrite reductase), and *narB* (nitrate reductase) have previously been used as indicators of low nitrogen stress [107, 108]. Each of these genes showed elevated expression in pro15 and syn12 (Fig 2.9B). Together the sample weights for these four clusters illustrate how *Prochlorococcus* and *Synechococcus* experience a progression from nitrogen scarcity in the subtropical gyre to iron scarcity in the subarctic gyre, consistent with estimates of nutrient limitation in the North Pacific surface ocean [60]. Moreover, the cluster profiles indicate a region of overlap in the subtropical-subarctic transition zone, where the sample weights for both the iron and nitrogen acclimation clusters showed elevated expression in both *Prochlorococcus* and *Synechococcus* populations (Fig 2.9C). Furthermore, in incubation studies of seawater collected from the transition zone, expression of the iron acclimation clusters was suppressed by supplementation with iron, and expression of the nitrogen acclimation clusters was suppressed by supplementation with nitrogen (Fig 2.9D). This simultaneous acclimation is distinctive to the transition zone; populations in the subtropical gyre were exclusively up-regulating the nitrogen acclimation clusters whereas populations in the subarctic gyre were solely up-regulating the iron acclimation clusters. Our data are consistent with previous reports of co-limitation at ecosystem boundaries [107, 109], and suggest that the transition zone may be a distinct ecosystem, shaped by multiple intersecting nutrient stresses, and inhabited by distinct populations equipped to cope with these conditions. Furthermore, the collective results of this study demonstrate the power of unsupervised signal discovery techniques to uncover processes that drive in situ microbiome dynamics, and draw a direct connecting line between the functions of individual genes, mechanisms of physiological response to environmental stressors, and ecosystem scale processes that mediate global biogeochemical cycles.

2.4 Methods

The code and data necessary to reproduce the analyses and evaluate the conclusions in the chapter are available as part of the publication Blaskowski et al. (2024) [52]. Supplementary Data files are available in an associated Zenodo repository: https://doi.org/10.5281/zenodo.12210994. The sequences reported have been deposited in the NCBI Sequence Read Archive under the following BioProject accession numbers: PRJNA1088528, PRJNA1090042, PRJNA1090086, PRJNA1090467, PRJNA1090899, PRJNA1091352, PRJNA492143, and PRJNA816919. Source code and documentation for the Barnacle model can be found in the package repository: https://github.com/blasks/barnacle. All code required to process data, run analyses and produce figures presented in this chapter can be found in an associated manuscript repository: https://github.com/blasks/barnacle-manuscript.

2.4.1 Metatranscriptomic data

In 2016, 2017, and 2019 a collaboration of ocean researchers collected samples for whole community metatranscriptomes from along a North Pacific transect ranging from around 22°N to 42°N along the 158th meridian west (Fig 2.1). Replicate seawater samples were pre-filtered through either 200 micron nylon mesh (2016 samples) or 100 micron nylon mesh (2017, 2019 samples) and then filtered onto 142 millimeter polycarbonate filters and immediately flash frozen in liquid nitrogen. Samples from surface transects of all three cruises, along with the 2017 depth profiles and 20 liter incubation samples were size fractionated by serial filtration onto 3 micron and 0.2 micron polycarbonate filters, whereas the samples from the 4 liter incubations in 2017 and the 2019 depth profiles and diel study samples consisted of a single size fraction collected on 0.2 micron polycarbonate filters. Whole community RNA was extracted from each filter using the Invitrogen ToTALLY RNA extraction kit (2016 samples) and the Zymo Direct-zol RNA Miniprep Plus kit (2017, 2019 samples). To focus sequencing libraries on messenger RNA, ribosomal RNA was depleted from extractions using either the Illumina Ribo-Zero Plus kit (2016, 2017 surface samples, depth profiles, and 4 liter incubations) and the ThermoFisher RiboMinus Transcriptome Isolation kit (2019, 2017 20 liter incubations). Libraries were sequenced at the University of Washington Northwest Genomics Center using an Illumina NextSeq 500 platform (2016 samples) and an Illumina NovaSeq 6000 platform (2017, 2019 samples). We quality controlled paired sequencing reads using Trimmomatic (version 0.32) [110], and combined reads from size fractionated samples, resulting in 222 samples of whole community RNA sequence data. Details of sample collection, RNA extraction, library preparation, and sequencing are available under the BioProject accession number associated with each dataset on the NCBI Sequence Read Archive.

We mapped sequencing reads against the 681 *Prochlorococcus* and *Synechococcus* reference genomes used

to generate the Cyanobacterial Cluster of Orthologous Group of proteins (CyCOG version 6.0; data S7 and S8) [62], using the read mapping software Salmon (version 1.10.2) [111]. We verified clade assignments for these reference genomes using a phylogeny of 424 single copy core CyCOGs that included additional genomes [112, 113, 114]. Single copy core CyCOGs were defined as those observed once in all 92 *Prochlorococcus* and *Synechococcus* genomes (> 99% complete) that were derived from cultivated isolates in CyCOG version 6.0. We used blastp (NCBI blast suite version 2.14.0) to identify CyCOGs in each additional publicly available genome. Each CyCOG protein family was aligned using clustal omega (clustalo version 1.2.4) [115], and CyCOG protein alignments were concatenated into a single multi fasta file using MEGA (version 11.0.13) [116]. Phylogenies for the *Prochlorococcus* and *Synechococcus* concatenated protein alignments were inferred separately in FastTree (version 2.1.11) [117] using the LG model of evolution with 100 bootstraps. We then assigned clade labels to the 681 reference genomes from CyCOG version 6.0 based on monophyletic group membership. One *Synechococcus* clade included reference genomes for strains CC9616 and KORDI-100 that have alternately been described as either clade UC-A or EnvC, which we delineate as UC-A for simplicity. Following mapping, we retained only the read counts corresponding to reference sequences annotated with both taxonomic clade and version 6.0 CyCOGs. Read counts were then aggregated by clade so that transcript abundances were consolidated to 24 pan-transcriptomes (8 *Prochlorococcus* and 16 *Synechococcus*). We restricted analysis to clades with at least one sample that recruited reads such that the percentage of detected pangenome CyCOGs surpassed 40% for *Prochlorococcus* or 60% for *Synechococcus*, corresponding to roughly 2,000 detected CyCOGs. The following clades met this threshold and were retained for further analysis: *Prochlorococcus* HLI, HLII, and LLI, and *Synechococcus* I, II, III, IV, CRD1, CRD2, and UC-A.

2.4.2 Normalization

To disentangle transcript abundance from organism abundance and focus our analysis on the most differentially expressed genes, we used version 2 of the variance stabilizing transform (vst) [64] to normalize the mapped read count matrix of each clade. Briefly, the vst fits a separate generalized linear model to the transcript abundance counts of each gene, using a negative binomial distribution as the link function [63], which accommodates the overdispersed mean-variance relationship observed in raw transcript abundance profiles (Fig 2.4). In each sample the transcript abundance of the gene is modeled as a function of overall clade abundance, approximated by the the total number of reads recruited to the clade pangenome in that sample. Regression parameters are then regularized across all gene models within each clade pangenome. Model residuals are output as the normalization product, indicating the degree and direction to which the transcript reads detected in each sample diverge from the expectation of the vst model. Before applying the normalization method, we pre-processed the read count matrix of each clade with detection and coverage

thresholds. These thresholds retained genes detected in at least 3 samples, and samples in which the percentage of detected genes achieved at least 1% of the maximum per-sample coverage observed in the clade pangenome. To account for dataset-scale bias in transcript counts that arise from technical artifacts (e.g. discrepancies in sample processing, library preparation, sequencing platform, etc.) we passed the sequencing batch to the vst function to be regressed out as a nuisance variable. All other vst arguments were set to the version 2 defaults, except 'min_cells', which was set to match the 3 sample detection threshold. This procedure produced for each clade a matrix of normalized residual transcript abundance values in which variance was de-correlated from mean abundance.

2.4.3 Data tensorization

The normalized residual transcript abundance matrices were segregated by genus and arranged into two gene (CyCOG) × taxon (clade) × sample tensors by aligning CyCOGs between clades. Any CyCOGs or samples not detected in a particular clade were filled in with zero values. The resulting *Prochlorococcus* data tensor encompassed 5,084 CyCOGs, 3 clades (HLI, HLII, and LLI) and 178 samples. The *Synechococcus* data tensor encompassed 6,161 CyCOGS, 7 clades (I, II, III, IV, CRD1, CRD2, and UC-A), and 222 samples.

2.4.4 Model fitting

All instantiations of sparse tensor decomposition models were fit using the 'SparseCP' API of the Barnacle package. We oriented data tensors gene × taxon × sample so that the l1 sparsity penalty was applied to the first mode, and non-negativity and unit l2 norm constraints were applied to the second and third modes. In cases where the sparsity coefficient was set to $\lambda = 0$, the unit l2 norm constraint was removed from all modes. We considered the algorithm to have converged when the change in loss dropped below 10^{-5}. We repeated all decompositions with five different random initializations, and the solution corresponding to the lowest loss was saved for analysis. All decompositions performed in this study converged in under 2,000 iterations. The random state of each decomposition was initialized with a unique integer seed, and to ensure reproducibility the seed of each saved model was stored as a local text file alongside model solutions and parameters.

2.4.5 Parameter selection

To select appropriate values of R and λ, we utilized a cross-validated grid search strategy that made use of sample replicates in the metatranscriptomic datasets. Data tensors were split by replicate along the sample axis to produce three replicate subtensors of shape $I \times J \times K_A$, $I \times J \times K_B$ and $I \times J \times K_C$, where I is the number of genes in the full dataset, J is the number of taxa in the full dataset, and K_A, K_B and K_C are the

number of samples in replicate set A, B and C, respectively. We then fit a series of models to each replicate subtensor using a grid search of different R and λ parameter values. Six cross-validated SSE scores were calculated for each unique set of parameters by comparing each fit model against the two held out replicate subtensors. Three cross-validated FMS scores were calculated for each parameter set by comparing the components between each pair of replicate models. In fitting both the *Prochlorococcus* and *Synechococcus* data tensors, we first pursued a coarse parameter grid search of $R \in [1, 5, 10, 15, 20, 25, 30, 35, 40, 45, 50]$ and $\lambda \in [0.0, 0.1, 1.0, 10.0, 100.0]$ in order to explore general trends. We examined the cross-validated SSE scores of the $\lambda = 0.0$ models to identify the the minimum error model in the absence of l1 regularization. The number of components was set to the R value corresponding to the this minimum. After a fine-tuned search of sparsity coefficients with $\lambda \in [1., 2., 3., ..., 31., 32., 64.]$ we then selected the sparsity coefficient as the maximum λ value at which the cross-validated FMS remained within one standard error of the maximum FMS.

2.4.6 Bootstrapping

For the *Prochlorococcus* and *Synechococcus* models, we used bootstrapping in conjunction with cross validation in order to mitigate bias originating from inconsistent numbers of replicates between samples (Fig 2.1), and to estimate the consistency of the resulting models. The 222 metatranscriptome samples originate from 87 unique conditions, of which 55 were sampled in triplicate, 25 in duplicate, and 7 were represented by a single sample. In each bootstrap, 'A', 'B', and 'C' replicate labels were randomly shuffled for each of the 55 conditions sampled in triplicate. For conditions with less than three replicates, replicate labels were randomly drawn without replacement. This ensured that the three resulting replicate subtensors were roughly equivalent in size, and that every sample had an equal and independent chance of being placed in any replicate set in any given bootstrap. Models were then fit independently to each replicate subtensor. Cross-validated SSE and FMS scores were calculated using only the indices of the samples common to the pair of replicate subtensors being compared. Consequently, conditions represented by a single replicate were used in fitting the models, but were not included in score calculations. For every unique set of parameters in the grid search, we fit 10 bootstraps, each with three replicate subtensors, for a total of 30 randomly shuffled data subtensors. For the best fit models, we fit 100 bootstraps of 3 replicate subtensors each, for a total of 300 shuffled replicate subtensors.

2.4.7 Robustness to mis-specification

Among bootstrap models parameterized with best fit parameters (as determined by cross-validated grid search), a "best representative" model was identified as the bootstrap with the highest mean FMS, calculated

in comparison to all other bootstrap models with best fit parameters. Models fit with alternate, suboptimal R and λ parameters were then aligned to this best representative model. Next, individual components were extracted from each aligned model and component-specific FMS scores were assessed for each component by comparing it against its matching component in the aligned best representative reference model. Additionally, the percentage of variation represented by each component of the best fit models was calculated using the 'percentage_variation' function of the 'tlviz' package [118].

2.4.8 Comparison of component weight profiles

For each component, we compiled a consensus weight profile from the collection of model bootstraps. Gene, taxon, and sample weight vectors were collected from each of 300 bootstrap replicates of best fit models, aligned against the best representative model. In the gene-mode, the median gene weight was calculated for each CyCOG, and the collection of CyCOGs with non-zero median weights was designated as a cluster. Additionally, the proportion of bootstraps in which the gene weights of a given CyCOG were non-zero was calculated as a measure of the strength of the association of a given CyCOG to a given cluster. The full set of *Prochlorococcus* and *Synechococcus* component weight profiles can be found in the Supplementary Materials published alongside Blaskowski et al. (2024) [52].

We compared gene-, taxon-, and sample-mode weight profiles between combined *Prochlorococcus* and *Synechococcus* components, considering only the most robust components with a cross-validated component FMS of 0.5 or higher. This threshold removed components pro1, pro14, syn1, syn2, and syn13 from further analysis. For each mode, a Pearson correlation was calculated between the median weight vectors of each pair of components, and the resulting correlation matrix was hierarchically clustered to identify groups of components with similar weight profiles. Additionally, a hierarchically clustered correlation matrix was calculated for a composite weight profile constructed by concatenating the median weight vectors of the gene-, taxon-, and sample- modes. To enable comparisons between *Prochlorococcus* and *Synechococcus* model components, any samples and CyCOGs specific to one model were added to the other model and filled in with zero values.

We also compared clusters from our analysis against previously identified clusters, published in a study that examined diel gene expression of *Prochlorococcus* strain MED4 grown over a light/dark cycle under nutrient replete laboratory conditions (Zinser et al., 2009) [69]. We used blastp (NCBI blast suite version 2.14.0) to identify the CyCOG associated with each MED4 gene, so that the CyCOG content of the Barnacle clusters could be directly compared to the CyCOG content of the Zinser clusters. Zinser et al. used a soft clustering algorithm to determined gene cluster membership, in which a score between 0 and 1 quantified the strength of the association of each gene to its cluster assignment, similar to the percent bootstraps

score calculated in this study. We used the 'f1_score' function of the 'sklearn.metrics' module to calculate a weighted F1 score between each pair of Zinser and Barnacle clusters, with weights set to the sum of Zinser cluster scores and Barnacle percent bootstrap scores. Null F1 scores were generated from randomly shuffled Zinser and Barnacle weight vectors, and an empirical cumulative density function (ECDF) constructed from the null scores was used to identify significantly similar cluster pairs, using a one-sided p-value threshold of 0.05 and Benjamini-Hochburg adjustment to control false discovery rate with multiple comparisons. We used the Python library 'statsmodels' to calculate the ECDF and adjust p-values, employing the 'ECDF' function of the 'statsmodels.distributions.empirical_distribution' module and the 'multipletests' function of the 'statsmodels.stats.multitest' module, respectively.

We compared component sample weight profiles against measurements of environmental conditions, accessed via the Simons Collaborative Marine Atlas Project (Simons CMAP) [119]. Estimates of surface chlorophyll were taken from the MODIS chlorophyll dataset in CMAP, an 8-day averaged product calculated using sea surface measurements from the MODIS Aqua satellite. The University of Southern California Marine Trace Elements Lab contributed dissolved iron measurements, and the David Karl lab at the University of Hawaii – Manoa contributed dissolved nitrate plus nitrite measurements.

2.4.9 Inference of circadian expression peak

We used a weighted kernel density estimate (KDE) to estimate the peak expression time of each component profile, assuming a circadian expression pattern. Across the combined metatranscriptome datasets, sampling times were not uniformly distributed around the 24 hour cycle. To account for uneven sample times, we fit a baseline KDE to the sample times of each of the *Prochlorococcus* and *Synechococcus* datasets, assigning equal weight to every sample. Then for each cluster, we fit a second weighted KDE to the component sample times, using as weights the product of the median sample-mode weight profile and the inverse of the appropriate baseline KDE. In effect, this biases the KDE towards samples in which the component signal is highly expressed while counteracting the KDE's inherent bias towards densely sampled times of day. The maximum of the weighted KDE was taken as the estimated peak expression time of each component. We used the 'gaussian_kde' function of the 'scipy.stats' Python library to fit KDEs.

2.4.10 Functional enrichment of gene clusters

We analyzed the biological function of each cluster by testing for over-representation of genes involved in common metabolic pathways or cell processes cataloged in the KEGG database (release 109.0). First, a consensus functional annotation was assigned to each CyCOG based on the best scoring annotation of a majority of member genes. Most CyCOGs had no annotation assigned to any member gene (36,693/40,295;

91.06%), but among the annotated CyCOGs, the majority had concordant functional annotations among member genes (3,474/3,602; 96.44%). Next, the majority annotations were used to assign CyCOGs to a subset of KEGG pathways that we manually curated to remove irrelevant processes, such as human disease pathways. Finally, the enrichment of each KEGG pathway was evaluated for each Barnacle cluster using a one-sided Mann-Whitney U test, as implemented by the 'mannwhitneyu' function of the 'scipy.stats' library. For each cluster, the test compares the gene weights of CyCOGs belonging to a particular pathway to the gene weights of CyCOGs not in that pathway, and evaluates the null hypothesis that the distribution of pathway gene weights is less than or equal to the distribution of non-pathway gene weights. In effect, this identifies the pathways for which member CyCOGs are significantly over-represented in the cluster, as compared to their background frequency in the expressed pangenome. We delineated enriched KEGG pathways as those with a p-value less than 0.01, following Benjamini-Hochburg adjustment for multiple comparisons using the 'statsmodels.stats.multitest' library. Consensus CyCOG annotations and KEGG pathway enrichments for each cluster can be found in the Supplementary Materials published alongside Blaskowski et al. (2024) [52].

3 Molecular signatures of antiphage activity in open ocean cyanobacteria

3.1 Introduction

Each liter of surface ocean seawater contains, on average, more viruses than there are humans on planet Earth [120]. A majority of these marine viruses are phages that target prokaryotic cells as hosts, including cyanophages that infect the abundant cyanobacterial photoautotrophs *Prochlorococcus* and *Synechococcus*. Cyanophages affect marine microbiomes in several significant ways, including as mortality agents that act as a biological control on cyanobacterial primary production, and as agents of genetic diversification [121]. Marine viruses such as cyanophages also affect biogeochemical cycles by means of the "viral shunt"; a process by which organic carbon and nutrients are released as dissolved organic matter as a result of viral infection and cell lysis [122]. Most of the dissolved organic matter released from cell lysis is subsequently recycled through the marine microbiome. In the North Pacific, a recently discovered hotspot of cyanophage abundance and infection was estimated to release enough viral lysate to supply up to 33% of the carbon demand of the local bacterial community [123]. Additionally, during infection cyanophages hijack cell metabolism in ways that are consequential for carbon and nutrient cycles, including by decreasing carbon fixation and redirecting energy derived from photosynthesis to nuceleotide biosynthesis [124]. Thus, the activities of the ocean's smallest organisms aggregate up to impacts that are global in scale. Therefore, characterizing the nature of biological interactions between cyanophages and their cyanobacterial hosts is critical to understanding marine ecosystem dynamics and biogeochemical cycles.

From an evolutionary perspective, phage predation presents a significant selection pressure on cyanobacteria populations. In the North Pacific, field measurements of active infections indicated that between 0.35% and 9.5% of *Prochlorococcus* cells are killed by viruses per day [125, 123]. As a result of this high selection pressure, cyanobacteria have evolved mechanisms to evade viruses and thwart infection. The exterior surface of the cell wall is the first line of defense against viral predation, and genes involved in determining the structure of the cyanobacterial cell wall have been implicated in resistance to different virus strains [126, 127, 128]. Additionally, like other prokaryotes, cyanobacteria have evolved molecular mechanisms designed to defend against viral attack. A recent survey of different environmental microbiomes found that nucleotide-sensing mechanisms such as restriction-modification systems, which discriminate and destroy foreign DNA, are widespread in marine ecosystems, and many have been detected in cyanobacterial genomes [129]. The same study also found genes in marine metagenomes that encode abortive infection systems, in which infected cells self destruct in order to prevent the spread of the virus to the rest of the susceptible

population. CRISPR-Cas systems of acquired immunity are apparently rare in marine cyanobacteria; few have been found in *Synechococcus* isolate genomes, and to date none have been reported in *Prochlorococcus* [130]. So far no known antiphage defense system is widespread in marine cyanobacteria. However, instances of arrested infection as a result of unknown mechanisms have been reported in laboratory studies of marine cyanobacteria [131], and despite the rapid pace of discovery of new antiphage mechanisms [132, 133], it is likely that the complete set of cyanobacterial antiviral defenses is not yet fully known.

Both as a result of selection pressure and because their lifestyle entails transferring foreign genetic material between cells, phage are understood to be significant drivers of genetic and phenotypic diversity in marine cyanobacteria populations. Beyond classically delineated *Prochlorococcus* ecotypes (HLI, LLII, etc.), one recent study from the North Pacific estimated that each milliliter of surface seawater contains thousands of genetically and phenotypcially distinct *Prochlorococcus* subpopulations [134]. This population diversity has been described as a "federation" of cells, unified by a common backbone of core genes, and distinguished by unique sets of flexible genes that fine-tune the adaptation of different subpopulations to different co-occurring niches [23]. Much of the flexible gene content of cells is concentrated in genomic island regions of elevated sequence variation [135], which has recently been linked to a newly discovered class of mobile genetic elements called tycheposons that seem to be specific to marine cyanobacteria [136]. Viruses are also understood to be drivers of horizontal gene transfer in cyanobacteria, as many cyanophage carry auxiliary metabolic genes that are homologous to core and flexible cyanobacterial genes. Perhaps an even more significant factor, under the Constant Diversity theory, diversity in natural prokaryotic populations is directly attributed to phage predation pressure [137]. In ecosystems such as the marine environment in which phage infection is a constant, the resulting selection pressure prevents any one clonal lineage from dominating the population, as these successful lineages would be subject to increased infection levels under "kill-the-winner" dynamics. The result is many subpopulations variously adapted to different micro-niches, each with its own pattern of resistance and susceptibility to different phage predators.

In this study we investigate mechanisms of cyanobacterial resistance and response to virus infection by examining the in situ gene expression patterns of *Prochlorococcus* and *Synechococcus* populations in the North Pacific. We focus on a cluster of co-expressed genes that showed elevated expression that coincided with the virus hotspot. We further characterize the functional profile of these genes by integrating data from previously conducted laboratory investigations on cyanobacteria facing phage infection. We consider the role of horizontal gene transfer in mediating phage-host interactions by examining the phylogeny and genomic context of genes linked to antiphage resistance. In detailing the in situ response of a cyanobacterial population to phage predation, this study offers a deeper understanding of the role of marine phages in shaping microbial population dynamics and diversity.

3.2 Results

In the previous chapter I used Barnacle to apply sparse tensor decomposition to *Prochlorococcus* and *Synechococcus* metatranscriptomes from environmental samples in the North Pacific. In both datasets, at least a dozen robust gene co-expression clusters were uncovered, each characterized by distinct spatiotemporal expression patterns, condition specificity, and enrichment for particular physiological processes. We identified diel periodicity in ten of the modeled expression patterns, and explored signatures of nitrogen and iron limitation exhibited by four of these. The identification of fifteen additional robust components suggests that the ecology of the North Pacific cyanobacteria population is affected by additional pressures not yet explored in previous analyses. In this chapter I turn to one of the additional fifteen clusters that preliminary analysis suggests is involved in host resistance and response to viral infection.

3.2.1 A coexpression cluster linked to virus infection

Cluster pro2 encompassed a set of 100 gene families (CyCOGs) that exhibited a pattern of anomalous transcript abundance, which peaked within the transition zone between the North Pacific Subtropical Gyre (NPSG) in the south and the Subpolar Gyre in the north [52]. The transition zone coincides with a hotspot for cyanophage abundance and infection that is hypothesized to significantly impact regional carbon and nutrient cycling [123]. Our analysis uncovered twelve KEGG pathways enriched in the pro2 cluster, including biofilm formation, peptidases, bacterial motility proteins, two-component regulatory systems, and seven pathways related to carbohydrate metabolism, glycosylation, and glycan biosynthesis. Cell surface structural features such as glycans are implicated in bacterial resistance to viral infection as these are common targets for virus attachment [137, 126]. Thus, I hypothesize that the pro2 cluster may reflect the in situ response of the transition zone *Prochlorococcus* population to an increased risk of viral infection.

The model attributed the pro2 expression pattern specifically to the high light II (HLII) clade of *Prochlorococcus*, which makes up a minor fraction of the *Prochlorococcus* population in the transition zone (Fig 2.2). Because Barnacle is an unsupervised pattern discovery tool designed to pull out the strongest trends in a dataset, less prominent trends may be excluded due to multilinearity (tensor) and sparsity constraints. Additionally, genes related to viral infection and response are often found in both host and virus genomes, complicating taxonomic assignment. To confirm whether the pro2 cluster is clade-specific or more broadly taxonomically distributed, we used the mapped metatranscriptomic data to investigate expression of the pro2 cluster genes across the entire *Prochlorococcus* and *Synechococcus* community.

Metatranscriptomic reads that mapped to pro2 CyCOGs were recruited by reference sequences from a variety of cyanobacterial taxa, including both the HLI and HLII clades of *Prochlorococcus*, as well as

different *Synechococcus* clades (Fig 3.1A). Fifteen of the 100 pro2 CyCOGs contain sequences originating from viral reference genomes, which altogether recruited less than 0.2% of reads overall, and no more than 1% of reads at any one station. Total transcript abundance varied widely among CyCOGs in the pro2 cluster, so a z-score transformation was used to enable comparison of whole community expression profiles between CyCOGs (Fig 3.1A). The z-scored expression profiles were comparable between pro2 CyCOGs, with relatively higher expression south of 32°N - 34°N (Fig 3.1, B-D). In 2017 and 2019, a smaller secondary peak in expression was apparent between 35°N and 37°N (Fig 3.1, C and D). In all three years, the taxonomy of pro2 transcript reads shifted from majority HLII *Prochlorococcus* in the south to majority HLI between 30°N and 36°N, whereas the majority of expression in the northernmost stations was predominantly attributed to *Synechococcus* species (Fig 3.1, E-G). These taxonomic trends are in agreement with the composition of the cyanobacteria community in North Pacific surface waters (Fig 2.2) and suggest a common set of physiological processes driven primarily by environmental conditions rather than taxonomic composition of the community.

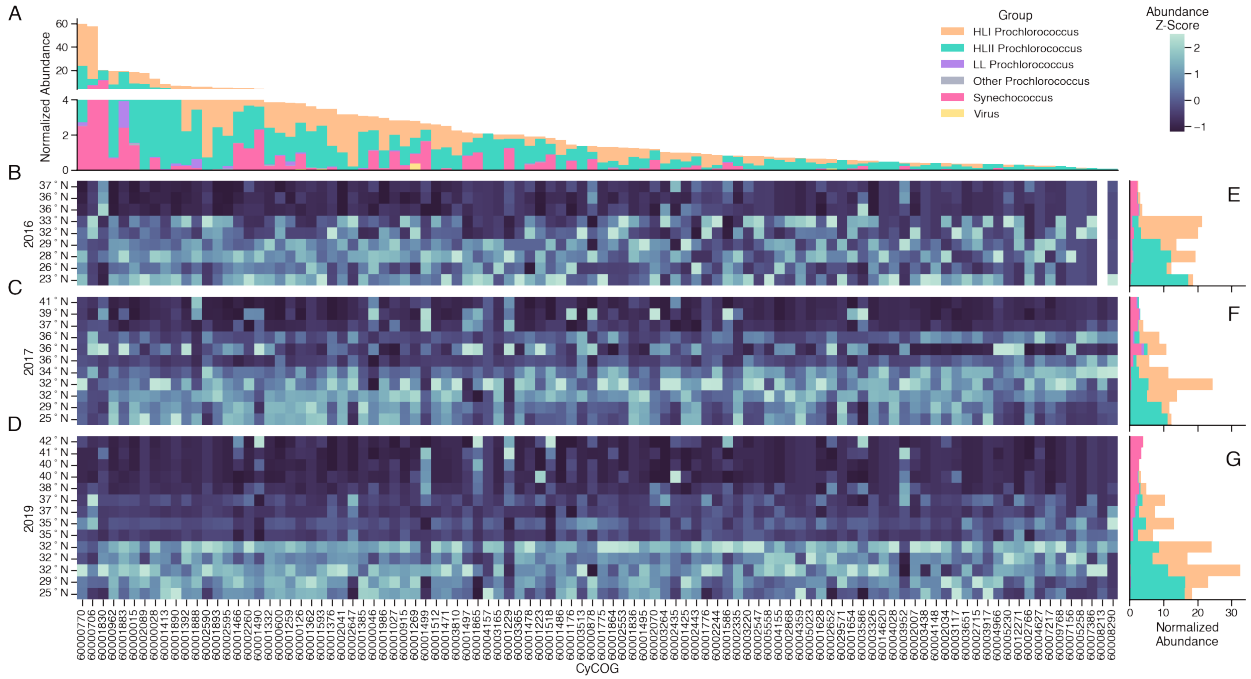

Figure 3.1: Cluster pro2 CyCOGs are transcribed by multiple *Prochlorococcus* and *Synechococcus* clades and exhibit similar latitudinal shifts in whole-community expression. (A) Total reads mapping to each CyCOG in pro2 cluster, colored by taxonomic group assignment. Read counts were normalized relative to single copy marker genes in each sample, then aggregated across the full dataset. Note discontinuous scale on y-axis. (B-D) Normalized whole-community transcript abundance of each CyCOG in each surface sample collected in (B) 2016, (C) 2017, and (D) 2019, averaged across replicates, and transformed column-wise by z-score to compare latitudinal profiles. (E-F) Total reads mapping to pro2 cluster CyCOGs in each sample, colored by taxonomic group assignment. Read counts were normalized relative to single copy marker genes, then aggregated across cluster CyCOGs.

I examined the spatial patterns of the pro2 expression profiles in more detail to evaluate their relationship to physical and ecological features of the North Pacific (Fig 3.2). In each of the three years, the average pro2 cluster expression profile was consistently elevated south of the salinity front, which defines the northernmost boundary of the NPSG. A decrease in pro2 expression occurred north of the salinity front boundary, coinciding with the transition zone between the NPSG and the Subpolar gyre, and the cyanophage hotspot identified by Carlson et al. (2022) [123] (Fig 3.2). Despite variation in the slope and smoothness of the decline, in all three years the hotspot aligned with a clear shift in the expression profile of the pro2 cluster, whereby the average expression level at every station south of the hotspot was markedly higher than the average expression level at every station north of the hotspot. This pattern, in combination with the enrichment for metabolic pathways involved in pathogen resistance and biosynthesis of cell surface structures, as well as the apparent taxonomic breadth of expression, motivated me to look for additional evidence linking the pro2 cluster CyCOGs to antiphage resistance and defense.

3.2.2 Evidence of antiphage resistance and defense

3.2.2.1 Antiphage resistance mutations

First, a number of mutations in specific *Prochlorococcus* and *Synechococcus* genes have previously been shown to confer resistance to viral infection in laboratory experiments [126, 127, 128]. We compiled data from these studies and identified a total of 54 CyCOGs that correspond to genes with at least one reported resistance-conferring mutation. Eight of these CyCOGs are included in the pro2 cluster (Table 3.1), constituting a significant ∼9-fold enrichment over expectation, given a baseline where the pro2 CyCOGs represent a random sample of the pangenome (Fisher's exact test, $p < 10^{-6}$). All eight CyCOGs are predicted to relate to cell surface structures. Five of the eight are involved in O-antigen biosynthesis, and a sixth is an under-characterized sugar-modifying enzyme (CyCOG 60001593). O-antigens are variable glycan structures that constitute the outermost layer of the lipopolysaccharide membrane that encapsulates many gram-negative bacteria, including marine cyanobacteria [138]. The two remaining CyCOGs are involved in macromolecular secretion, including a component of a family of ABC-transporters involved in bacterial Type 1 protein export (60001223; COG2274) and a small family of protein autotransporters (60003264; PFAM03797). Each of the identified CyCOGs encompasses multiple sequence variants, and further research is needed to determine which variants may confer resistance to different viral strains in wild populations. Nevertheless, these results implicate the pro2 cluster in antiphage resistance, and underscore the integral role of exterior cell surface structures in the susceptibility/resistance of cyanobacteria to viral infection.

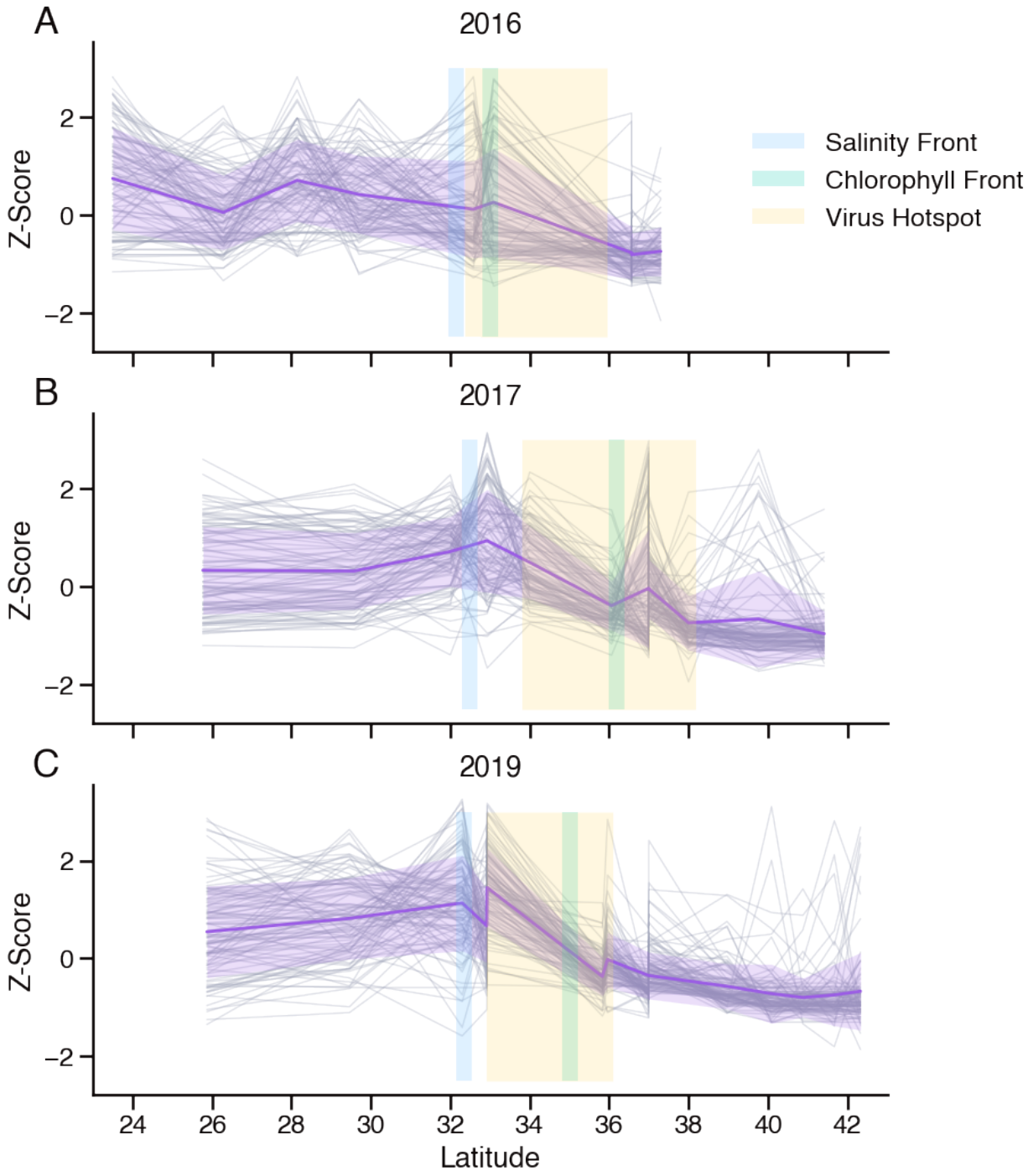

Figure 3.2: Expression level decrease in pro2 cluster profile coincides with virus hotspot. Z-scored whole community transcript abundance profiles (*Prochlorococcus* + *Synechococcus*) as a function of latitude, with each grey line representing a distinct pro2 cluster CyCOG, purple line indicating mean expression profile of the cluster, and purple band delineating standard deviation. Dataset consists of surface samples (< 15 m depth) collected in (A) 2016, (B) 2017, and (C) 2019. Salinity and chlorophyll fronts as reported by Juranek et al. (2020) [77] are indicated by the blue and green bars, and virus hotspot identified by Carlson et al. (2022) [123] is indicated by yellow bands.

Table 3.1: CyCOGs from pro2 cluster for which mutant homologs in *Prochlorococcus* or *Synechococcus* have previously been reported to confer to resistance to viral infection in a laboratory setting.

CyCOG	Description	Mutations	Strains	References
60000015	dTDP-4-amino-4,6-dideoxygalactose transaminase	11	MED4	[126, 128]
60000046	UDP-glucuronate 4-epimerase	6	MED4	[128]
60000878	UDP-galactose 4-epimerase	7	MED4	[126, 128]
60001223	ATP-binding cassette, subfamily B	2	MED4, MIT9312	[128]
60001497	Glucose-1-phosphate thymidylyltransferase	1	WH7803	[127]
60001499	N-acetylneuraminate synthase	1	MIT9312	[126]
60001593	carbamoyltransferase	1	MED4	[126]
60003264	Autotransporter beta-domain-containing protein	17	MED4	[126, 128]

3.2.2.2 Antiphage defense genes

Second, most prokaryotic genomes encode molecular systems involved in active antiphage defense, including well-known mechanisms such as the CRISPR-Cas system of acquired phage immunity, as well as currently uncharacterized systems inferred from their proximity to established defense systems in genomic regions termed "defense islands" [132]. DefenseFinder is an annotation tool that maintains an updated database of defense system profiles, each consisting of a set of linked Hidden Markov Models (HMMs), which can be used to identify putative defense genes and systems in reference genomes [139]. I annotated all *Prochlorococcus* and *Synechococcus* reference genomes in our dataset with DefenseFinder, and obtained results at two levels of stringency. The more permissive annotation results include any gene that is a significant match to any defense gene HMM in the DefenseFinder database. The more stringent annotations only include complete systems in which all gene components necessary for functionality were found co-located in a genome. In the joint *Prochlorococcus* and *Synechococcus* pangenome, DefenseFinder annotated a total of 14,895 genes, corresponding to 145 unique CyCOGs present in the North Pacific metatranscriptomes. Of these, eight putative defense CyCOGs were included in the pro2 cluster (Table 3.2), comprising a significant 3-fold enrichment over expectation, given random sampling of the pro2 CyCOGs (Fisher's exact test, $p < 0.01$). Under the more stringent requirements that take into account genomic context, 51 CyCOGs were linked to complete defense systems for which all necessary components were co-located. Contrary to a previous finding [130], these complete systems include a number of putative CRISPR-Cas defense systems in *Prochlorococcus* reference genomes, although none were included in the pro2 cluster. Four CyCOGs linked to complete defense systems were included in the pro2 cluster, representing a subset of the eight putative defense CyCOGs found with more permissive annotations, and a similarly significant enrichment over expectation given random sampling (Fisher's exact test, $p < 0.02$).

Table 3.2: CyCOGs from pro2 cluster that correspond to proteins putatively involved in antiphage defense systems.

CyCOG	# Proteins	Description	Defense System	# HMM Hits	# Complete Systems
60001376	1002	Tetratricopeptide repeat-containing protein	Lamassu-Fam	9	0
60002070	127	adenylate cyclase	Pycsar	111	0
60002435	75	type I restriction enzyme M protein	RM Type I	66	38
60002766	59	Helicase conserved C-terminal domain-containing protein	RM Type IIG	7	3
60003117	43	DNA recombination protein RmuC	Eleos	2	0
60003220	36	type I restriction enzyme, R subunit	RM Type I	25	21
60003326	34	Superfamily I DNA and/or RNA helicase	Mokosh type I	13	0
60007217	5	ATP-dependent Lon protease	BREX I	4	2

The eight pro2 cluster CyCOGs annotated as putative defense genes fell into six categories of defense systems. Three of the pro2 CyCOGs, all linked to complete defense systems, are components of restriction-modification systems (Table 3.2). These CyCOGs include a gene family encoding methyltransferases (CyCOG 60002435), which methylate host DNA with a unique pattern designed to enable self-recognition, and a family of restriction enzymes (CyCOG 60003220) which digest foreign DNA lacking the methylation pattern [140]. The fourth pro2 CyCOG, also linked to a complete defense system, encodes a component of the BREX antiphage defense system, which also uses methylation to distinguish foreign DNA, but employs an unknown mechanism other than DNA cleavage to prevent phage replication [141]. The remaining four putative defense CyCOGs in pro2 were not co-located with the complete suite of genes necessary for functionality in a previously validated defense system, but a subset of member proteins showed homology to components of known antiphage defense systems. These included components of Lamassu and Pycsar systems, which employ different mechanisms of abortive infection, in which a host initiates cell death in order to halt infection and protect the rest of the population [140]. The mechanism of defense remains uncharacterized for the Eleos and Mokosh systems that corresponded to the remaining two CyCOGs. The fact that pro2 encompasses CyCOGs corresponding to both components of a complete restriction-modification system points to active expression of antiphage defense systems by the North Pacific cyanobacterial population, and the inclusion of additional putative defense proteins in the cluster suggests a broader program of antiphage defense that remains to be fully characterized.

3.2.2.3 Differential expression in culture infection studies

As a third approach, we cross-checked the pro2 CyCOGs against a compiled dataset of cyanobacterial genes previously reported to exhibit significantly altered expression when exposed to viral infection in culture [142, 143, 144, 145]. This transcriptional response may include mechanisms of antiphage resistance and defense, as well as signatures of viral hijacking of host metabolism. In total, we found 188 CyCOGs that corresponded to genes with altered expression patterns in culture, of which 9 were included in the pro2 cluster (Table 3.3). This represents a greater than twofold enrichment over expectation under a random distribution (Fisher's exact test, $p < 0.02$). Three of the nine CyCOGs showed increased expression in culture in response to viral exposure, including one related to O-antigen biosynthesis (CyCOG 60000046) that also conferred resistance to viral infection when mutated (Table 3.1), and another (CyCOG 60003220) that is the restriction enzyme component of a restriction-modification defense system (Table 3.2). The remaining six CyCOGs showed unchanged expression levels throughout viral exposure experiments, which is notable since the majority of transcript levels decreased over the course of viral infection [143]. While none of these six CyCOGs are well-characterized, one is a family of nuclease homologs (CyCOG 60001628), two contain conserved domains often found in DNA repair proteins (CyCOGs 60002007 and 60002715), and a fourth contains a conserved domain linked to secretion pathways (CyCOG 60001890). Collectively these findings show a pattern of elevated expression correlated with elevated viral abundance that is detected in both the field and lab, which seems to be related to the synthesis and export of cell surface structures, as well as nucleotide degradation and repair, perhaps as part of a broader host antiphage defense response.

Table 3.3: CyCOGs from pro2 cluster that correspond to genes that exhibited differential expression in laboratory culture experiments after exposure to viral infection.

CyCOG	Description	Strain	Phage	Expression Response	Reference
60000046	UDP-glucuronate 4-epimerase	MED4	P-SSP7	increased	[142]
60001628	nuclease homolog	WH8102	Syn9	unchanged	[143]
60001830	S-layer giant protein SwmA	WH8102	Syn9	unchanged	[143]
60001864	hypothetical protein	WH8102	Syn9	unchanged	[143]
60001890	prepilin-type N-terminal cleavage/methylation domain-containing protein	WH7803	Syn9	unchanged	[143]
60002007	AAA domain-containing protein	WH8102	Syn9	increased	[143]
60002034	Virulence-associated protein E	WH8102	Syn9	unchanged	[143]
60002715	AAA domain-containing protein	WH8102, WH8109	Syn9	unchanged	[143]
60003220	type I restriction enzyme, R subunit	WH8102	Syn9	increased	[143]

3.2.2.4 Summary of evidence

In summary, by integrating evidence from three independent data sources, we found a remarkably cohesive functional profile for cluster pro2, with some level of support for antiphage activity for 70% (46/66) of the functionally annotated CyCOGs (Fig 3.3). Experimental evidence points to roles in host antiphage resistance, defense, or response to virus exposure for 23 of the 100 CyCOGs in cluster pro2. Evidence from functional annotations links an additional 23 of the pro2 CyCOGs to the composition and structure of the cell surface. The annotations fall into similar pathways and processes found in the experimental data; nine of the 23 surface structure CyCOGs are connected to lipopolysaccharide and O-antigen biosynthesis pathways, four are involved in polysaccharide and polypeptide export, and the remaining ten are comprised of transporters and membrane-bound proteins with varying levels of characterization. Three of the less-characterized CyCOGs (60001830, 60001883, and 60001888) appear to be families of giant proteins, one of which (CyCOG 60001830) includes the 140 kDa protein subunit SwmA that assembles in many copies to form a geometric glycoprotein S-layer (surface layer) exterior to the outer membrane in a strain of marine *Synechococcus* [146]. Other giant proteins co-localize with the S-layer and form different cell surface structures, which are hypothesized to perform functions ranging from cell motility to avoiding predation by viruses and grazers [57]. Of the remaining 54 cluster CyCOGs, the protein function is unknown for 34. Altogether the experimental and annotation presents a clear picture of antiphage activity in the pro2 cluster, and points to a similar functional role for the co-expressed CyCOGs that are not yet characterized.

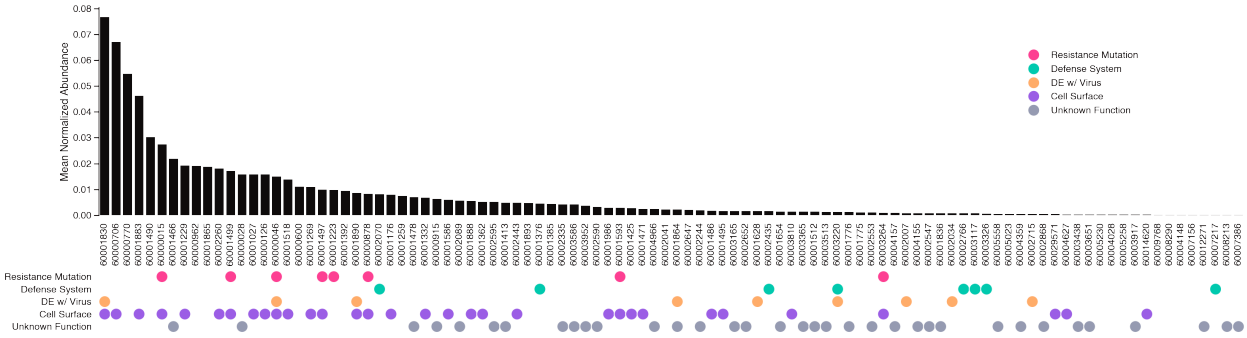

Figure 3.3: Nearly half of pro2 cluster CyCOGs show some evidence of involvement in host resistance or response to viral infection. Pink markers indicate CyCOGs corresponding to proteins that confer resistance to viral infection when mutated; green, putative antiphage defense system proteins; orange, genes that showed differential expression in culture in response to virus exposure; and grey, proteins of unknown function. Bars indicate the geometric mean of normalized whole-community transcript abundance in North Pacific surface metatranscriptomes.

3.2.3 Antiphage genes in genomic islands

In *Prochlorococcus*, genetic variation is not uniformly distributed throughout the physical genome, but rather is concentrated in "genomic islands" characterized by high sequence diversity [135], partially as a result of increased levels of horizontal gene transfer [136]. As a result, core genes that are common to all *Prochlorococcus* lineages are more often positioned in conserved genomic regions outside genomic islands, whereas flexible genes, found only a subset of lineages, tend to concentrate in genomic islands. Genes related to viral resistance and defense have also been found at high concentrations in genomic islands [126]. To investigate the proximity of each CyCOG to a genomic island, we utilized a dataset of *Prochlorococcus* genomes for which genomic island locations had been predicted using a custom EM-type algorithm [136]. We matched each gene in the Hackl et al. dataset to its corresponding CyCOG, and designated core CyCOGs as those with at least one ortholog detected in every isolate genome in the dataset and flexible CyCOGs as those not universally detected. In the resulting dataset we found that the median genomic island spans 27,949 base pairs, and that the median gene corresponding to a flexible CyCOG is positioned 2,070 base pairs within a genomic island. In contrast, the median gene corresponding to a core CyCOG is positioned 46,482 base pairs outside of a genomic island (Fig 3.4A). Aggregating proximities within each CyCOG, we found that for 75% of flexible CyCOGs the median ortholog sits within a genomic island, whereas for 95% of core CyCOGs the median ortholog is positioned outside of a genomic island. Furthermore, we found a significant positive correlation between the median distance from a genomic island, and a CyCOG's prevalence in isolate genomes (data not shown).

To determine whether the pro2 cluster CyCOGS are preferentially distributed within genomic islands, we summarized the median genomic island proximities for the pro2 CyCOGs and compared this distribution to that of all CyCOGs within *Prochlorococcus* pangenome (Fig 3.4, B and C). A summary proximity was determined for each CyCOG as the median distance to the nearest genomic island among member genes. The summary proximity of the median pangenome CyCOG (including both core and flexible CyCOGs) was 4,083 base pairs inside a genomic island, with 35% of CyCOGs predominantly positioned outside a genomic island (Fig 3.4B). In cluster pro2, the summary proximity of the median CyCOG was 7,584 base pairs within a genomic island, with just 9% of CyCOGs predominantly positioned outside (Fig 3.4C), representing a significant shift in proximity towards genomic islands when comparing the pro2 cluster CyCOGs to the pangenome as a whole (Mann-Whitney U test, $p < 10^{-5}$). This result reiterates the connection between genomic islands and genes involved in antiphage activity in cyanobacteria. However the trend is not absolute as the the pro2 cluster encompasses a mix of core and flexible genes found both within and outside genomic islands, leading us to next investigate the population and genomic distribution of each pro2 CyCOG in more

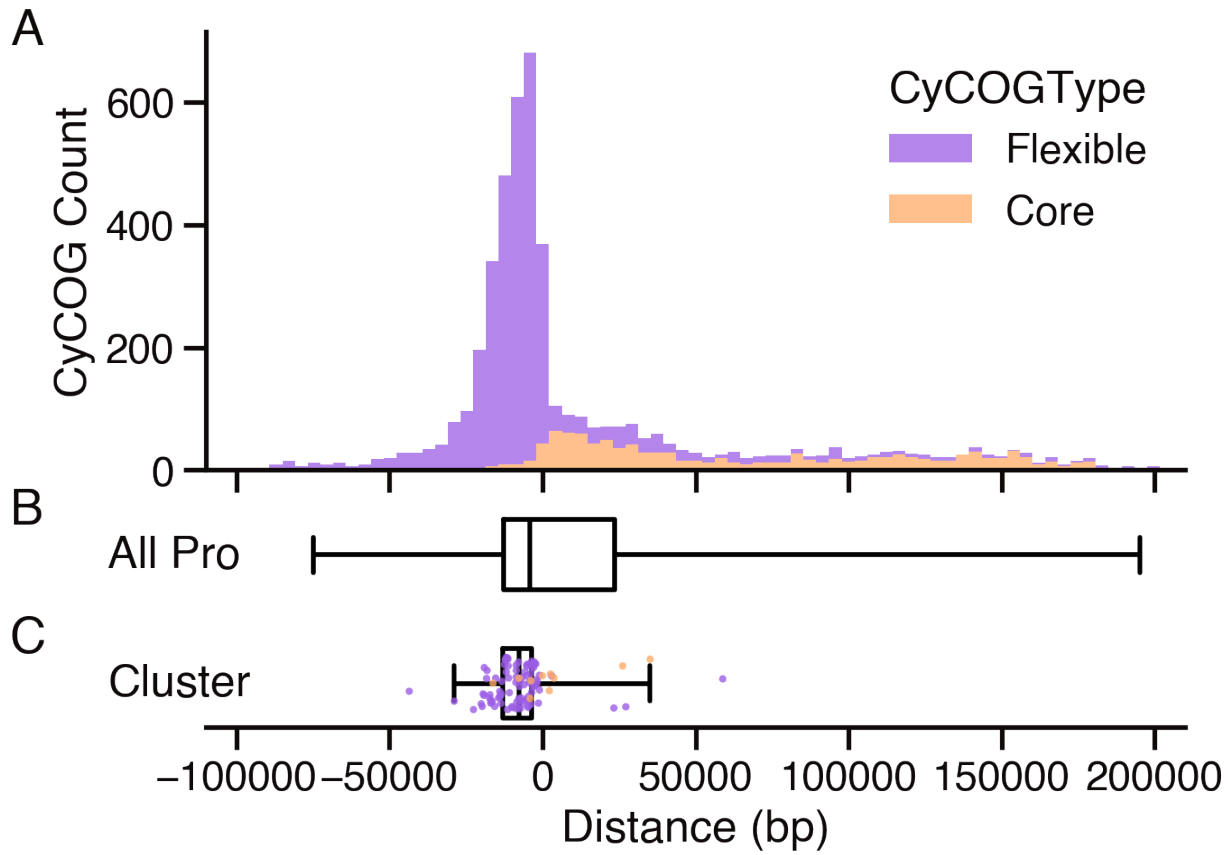

Figure 3.4: Cluster pro2 CyCOGs are biased towards positions within genomic islands, compared to the broader *Prochlorococcus* pangenome. (A) Histogram of median proximity to nearest genomic island for each CyCOG, broken down by core (orange) and flexible (purple) CyCOGs. (B) Distribution of median genomic island proximities across full pangenome. (C) Distribution of median genomic island proximities for pro2 cluster, with the value for each pro2 CyCOG plotted as individual points. Boxes indicate inner 50th percentile, centered on the median, and whiskers indicate 1st and 99th percentiles.

detail.

3.2.4 Antiphage genes in the core and flexible pangenome

Many genes that confer a fitness advantages to cyanobacteria, including some involved in resistance to viral infection, are present only in distinct subpopulations as part of the extensive flexible pangenome. However, such a binary categorization of core/flexible genes obscures a distributional gradient in the flexible pangenome that gives rise to the phenotypic heterogeneity of marine cyanobacterial populations [23]. I therefore determined the degree to which each CyCOG is shared between subpopulations by calculating the percentage of isolate *Prochlorococcus* genomes containing an ortholog of each CyCOG. This percentage served as a "coreness" metric, with a higher coreness indicating CyCOGs present in a higher proportion of reference genomes. There was a slight trend in which pro2 cluster CyCOGs with higher coreness showed a higher

average transcript abundance in our dataset, although many CyCOGs broke with this trend (Fig 3.5A). At a more granular taxonomic scale, we found several distinct patterns with regards to population distribution and genome copy number (Fig 3.5B). Ten of the pro2 CyCOGs were universally detected in all isolate *Prochlorococcus* genomes. These ten CyCOGs tended to be core to marine *Synechococcus* clades 5.1A and 5.1B as well, with the exception of CyCOGs 60000028 and 60000600. Three of the core pro2 CyCOGs (60000015, 60000028, 60000046) were found in multiple copies per genome, on average, while the remaining core CyCOGs appear to be single copy genes. It is noteworthy that two of the multicopy core CyCOGs have also been found in isolate cyanophage genomes. A second group of pro2 CyCOGs were commonly single copy and core to particular clades, but not universally distributed among *Prochlorococcus* and *Synechococcus* genomes. Interestingly, a subset of these CyCOGs are present in up to nine copies per genome in some subpopulations and include all three of the giant proteins potentially involved in forming the cell surface S-layer (60001830, 60001883, and 60001888). Finally, a long tail of rare CyCOGs are uncommonly detected in a small proportion of genomes, potentially because these CyCOGs confer more localized fitness advantages to the host strains.

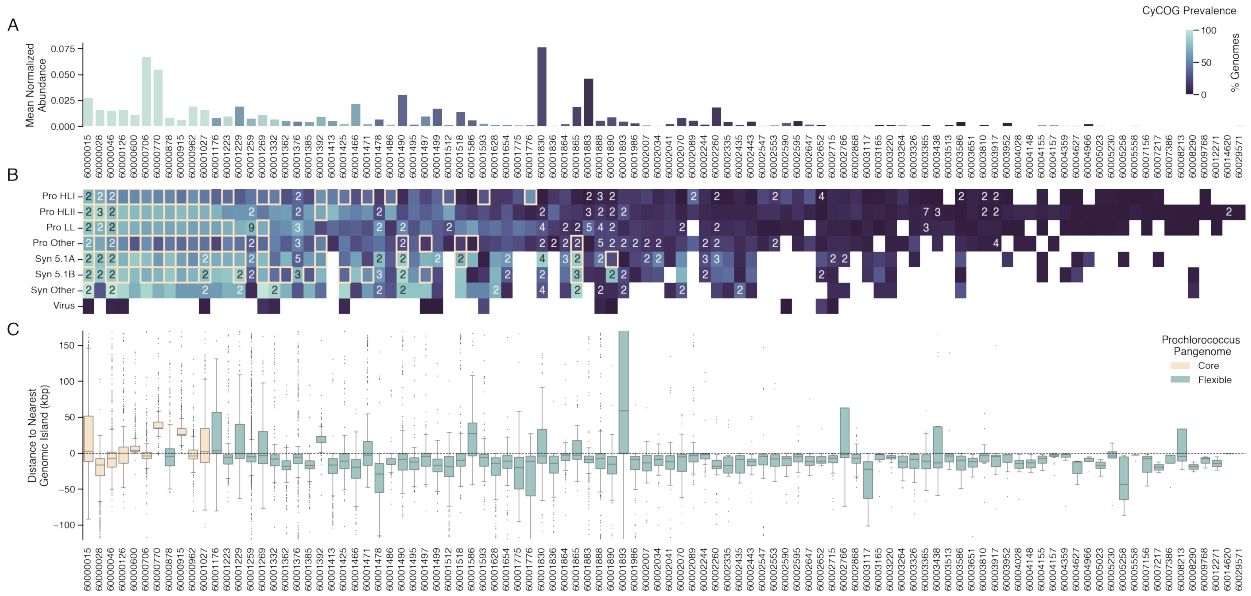

Figure 3.5: Cluster pro2 encompasses a mix of core and flexible CyCOGs with varied distributions across sequenced genomes and genomic islands. (A) Geometric mean of normalized transcript abundance of each pro2 CyCOG, colored by the percentage of *Prochlorococcus* reference genomes with a corresponding ortholog. (B) Prevalence of each pro2 CyCOG in reference genomes of disaggregated taxonomic groups. Outlined boxes indicate instances in which a CyCOG ortholog was found in 100% of isolate reference genomes of the corresponding group, and numbered boxes indicate the average genomic copy number of CyCOGs detected two or more times per reference genome. (C) Proximity of CyCOG proteins to nearest genomic island. Negative values indicate positions within genomic islands, and positive values positions outside. Plots show the distribution of distances across *Prochlorococcus* reference genomes, in which boxes indicate the interquartile range centered on the median, whiskers indicate 5th and 95th percentiles, and outlying points are plotted individually. Box plots are colored to distinguish flexible CyCOGs from core, defined as those found in 100% of isolate *Prochlorococcus* genomes.

We compared population distribution patterns with genomic location information by integrating the genomic island proximity data for each pro2 cluster CyCOG (Fig 3.5C). Overwhelmingly, the pro2 CyCOGs are situated within genomic islands, regardless of their prevalence in *Prochlorococcus* and *Synechococcus* populations. However, a fraction of both core and flexible CyCOGs were located primarily outside of genomic islands. In fact, the ten CyCOGs core to *Prochlorococcus* were evenly split, with five primarily situated within genomic islands, and the other five predominantly located outside genomic islands. This observation highlights two intriguing patterns: 1) core genes that are universally distributed among *Prochlorococcus* strains, but nevertheless subject to the elevated selection pressure and rates of horizontal gene transfer associated with genomic islands, and 2) core genes that are not found in genomic islands and thus are not subject to these evolutionary forces, but are nevertheless co-expressed with this suite of high-diversity genes. To explore these observations more deeply, we selected four core CyCOGs representing a range of proximities to genomic islands as case studies for investigating the evolutionary history of the member sequences.

3.2.5 Horizontal gene transfer of antiphage genes

We examined the phylogenies of orthologs from four core CyCOGs present in all *Prochlorococcus* reference genomes. Two of these CyCOGs – 60000770 and 60000915 – are both situated outside of genomic islands by a median distance of 35,189 and 26,206 base pairs, respectively. The other two CyCOGS – 60000028 and 60000046 – are positioned within genomic islands by a median of 16,159 and 7,725 base pairs, respectively (Fig 3.5C). For each of these four CyCOGs, we extracted all orthologs originating from isolate *Prochlorococcus* or *Synechococcus* genomes, and generated phylogenetic trees from the gene nucleotide sequences. The two CyCOGs found outside genomic islands produced gene phylogenies that largely agreed with the clade designations of their genomes of origin (Fig 3.6, A and B). In contrast, the two CyCOGs found within genomic islands had gene phylogenies that disagreed with the clades of their originating genomes (Fig 3.6, C and D). Furthermore, these latter phylogenies displayed more branching in comparison to those of the CyCOGs from outside genomic islands, indicating a higher degree of sequence diversity, perhaps as a result of greater diversifying selection pressure. These results are consistent with a pattern of vertical inheritance for genes located outside of genomic islands, and a pattern of horizontal gene transfer for genes located within the genomic island regions.

To look for a broader signature of horizontal gene transfer within the metatranscriptomic data, we compared the variance of each CyCOG's transcript abundance values across three levels of taxonomic granularity (Fig 3.7). We reasoned that CyCOGs subject to a high degree of horizontal gene transfer would exhibit elevated variance when normalized at a more granular taxonomic level as a result of mismatches between the taxonomy of the organism transcribing the gene in situ and the taxonomy of the reference sequence recruiting

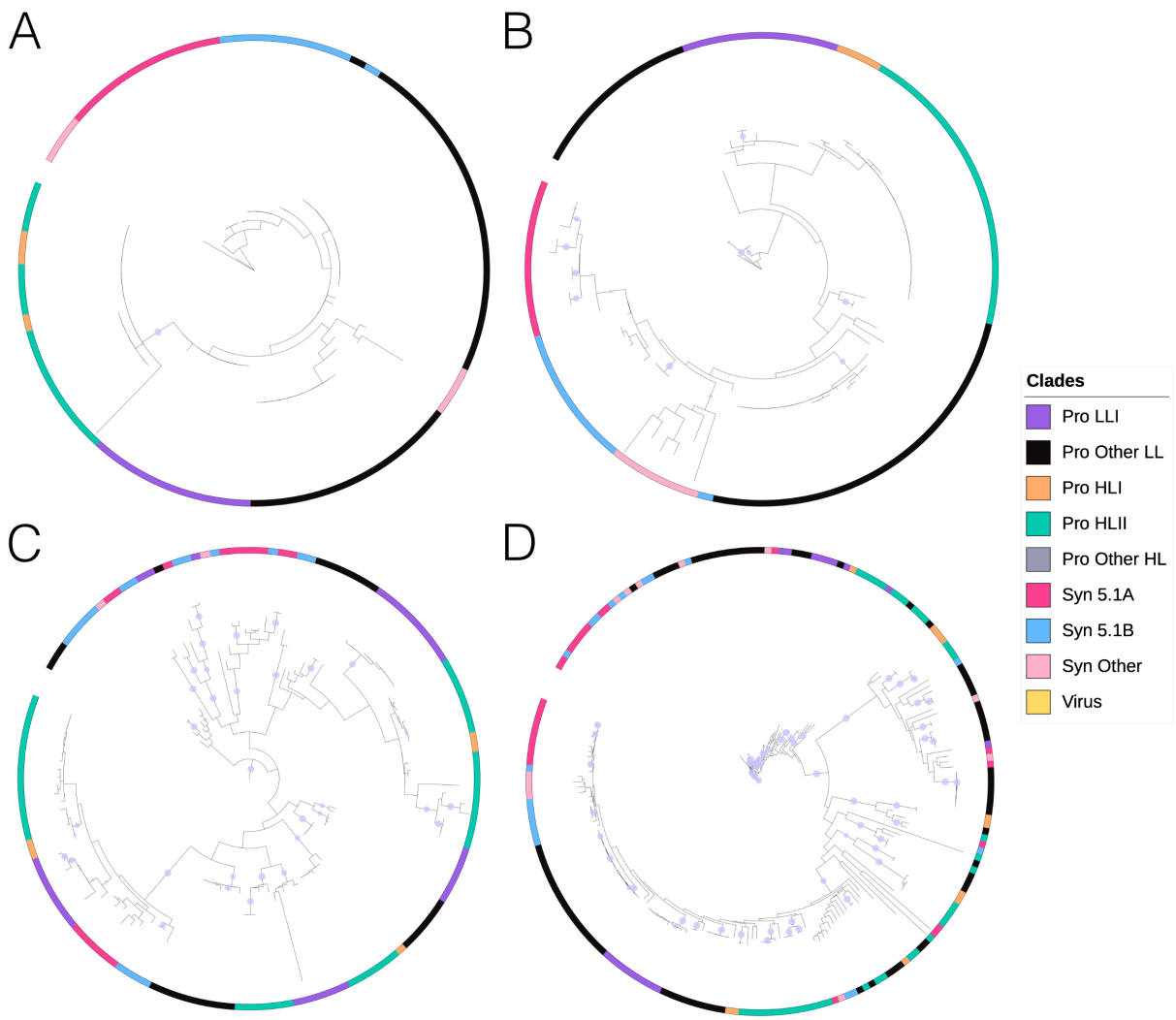

Figure 3.6: Phylogenies of core CyCOGs positioned within genomic islands are discordant with genome clade assignment. Un-rooted phylogenetic trees (amino acid sequence) of core CyCOGs located outside (A, B) and inside (C, D) of genomic islands. (A) CyCOG 60000770 – family of bacterial DNA binding protein, positioned a median of 35,189 base pairs outside the nearest genomic island. (B) CyCOG 60000915 – protein of unknown function (DUF1825), median 26,206 bp outside genomic island. (C) CyCOG 60000028 – protein of unknown function (DUF3764), median 16,159 bp inside genomic island. (D) CyCOG 60000046 – UDP-glucuronate 4-epimerase, median 7,725 bp inside genomic island. Leaf colors indicate clade assignment of reference genome of origin, and markers indicate branches supported by at least 95 of 100 bootstraps.

transcript reads during mapping. We found a general trend that most CyCOGs displayed higher variance at the more granular taxonomic level, for instance comparing normalization at the clade level (HLII) with normalization across the *Prochlorococcus* and *Synechococcus* community as a whole (Fig 3.7A). A similar trend was evident when comparing normalization at the genus level (*Prochlorococcus*) with whole community normalization as well, albeit with a smaller average shift (Fig 3.7B). However, a subset of flexible CyCOGs displayed a stronger variance shift than the rest of the pangenome under clade level normalization, which

was not apparent in the genus level normalization. The majority of the pro2 CyCOGs fall within this subset (Fig 3.7, A and B). We found that the majority of the subset of CyCOGs with elevated clade level variance are situated within genomic islands (Fig 3.7, C and D). Together with the phylogenetic case studies, these results indicate that the genes expressed in the pro2 cluster are characterized by a pattern of frequent horizontal gene transfer between clades, suggesting a diversifying selection pressure for these genes involved in virus resistance and defense.

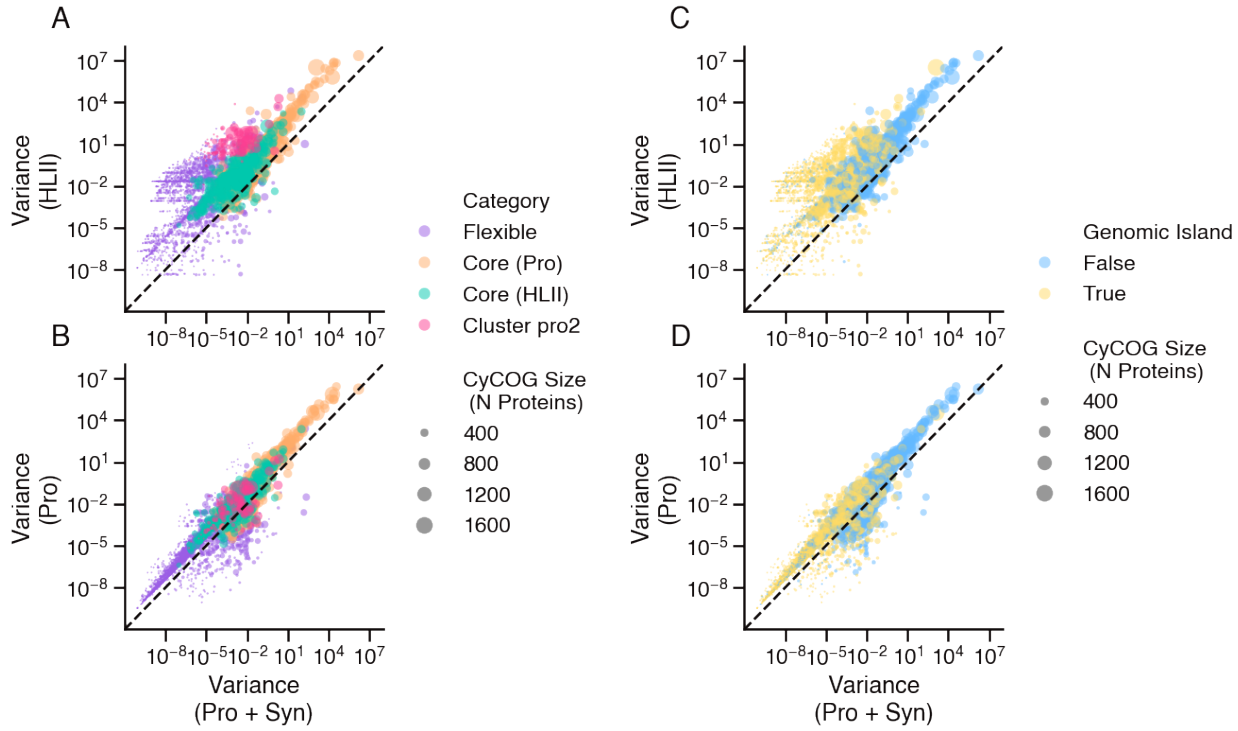

Figure 3.7: Normalization at the clade level as compared to the genus level results in elevated variance for flexible CyCOGs located in genomic islands. Scatter plots comparing variance of transcript abundances for (A, C) CyCOGs normalized at the clade level versus whole community normalization (*Prochlorococcus* + *Synechococcus*) and (B, D) CyCOGs normalized at the genus level (*Prochlorococcus*) versus whole community normalization. (A, B) Markers colored by population distribution and membership in the pro2 cluster. (C, D) Same plots as A and B, respectively, but colored by location within genomic islands. Markers are sized according to the number of protein sequences in the CyCOG group.

3.3 Discussion

Here we provide evidence that a co-expressed set of 100 gene families (CyCOGs) plays a role in antiphage resistance and defense in natural populations of *Prochlorococcus*. The co-expression cluster, referred to as pro2, is one of 25 robust co-expression clusters previously identified using tensor decomposition of marine metatranscriptomic data collected in the North Pacific [52]. The pattern of expression modeled by the pro2 cluster is shared across the dominant cyanobacterial taxa of the North Pacific, with expression detected in the

high light *Prochlorococcus* clades HLI and HLII, the low light clade LLI, and various clades of *Synechococcus* (Fig 3.1). As a collective, these cyanobacterial populations displayed higher relative expression of the pro2 cluster CyCOGs in the oligotrophic North Pacific Subtropical Gyre (NPSG), and lower expression in the transition zone between the NPSG and the Subpoloar Gyre (Fig 3.2), a region coinciding with a previously identified area of significantly elevated cyanophage abundance and infection [123]. Functional evidence for about 75% of the pro2 CyCOGs is consistent with roles in viral resistance and defense (Fig 3.3). Furthermore, a majority of the genes associated with the pro2 cluster are located within genomic islands (Fig 3.4), in agreement with previous findings that antiphage resistance and defense genes are concentrated in these regions [135, 126]. Altogether, these findings provide evidence for a population level response to viral predation, and identify potential mechanisms of antiphage resistance and defense common among natural populations of marine cyanobacteria.

Since viruses must first attach to a target cell in order to initiate infection, the exterior surface of the cell serves as the first line of defense against phage infection. The cell envelope of cyanobacteria is typical of gram-negative bacteria and consists of a plasma membrane surrounded by a peptidoglycan layer, followed by an outer membrane encased by a layer of lipopolysaccharides (LPS) [147]. As the outermost layer of the cell wall, LPS is a known target for viral attachment in cyanobacteria [148]. Additionally, in some cyanobacteria a surface layer (S-layer) mesh of glycoprotein sits on top the LPS layer of the outer membrane [149, 146]. In general, up to 75% of the cell surface area consists of LPS, with the rest made up of a mix of proteins, glycoproteins, and integral membrane transporters [150]. About a third of the pro2 cluster CyCOGs (32/100) appear to be involved in determining the composition and structure of the cell surface, falling into three general categories: enzymes involved in LPS biosynthesis, giant proteins related to the S-layer, and a variety of transporters, including some involved in translocating macromolecules to the outer membrane. Direct experimental evidence and gene functional annotations implicate members of each category in host antiphage resistance (Fig 3.3).

A total of 15 pro2 cluster CyCOGs are putatively involved in LPS biosynthesis, including epimerases, transferases, transaminases, and other glycosyl modification enzymes that determine the structure of the outermost O-antigen layer of LPS. Experimental evidence has previously shown that different sequence variants for six of these enzymes result in different patterns of resistance to cyanophage in culture (Table 3.1). For example, sequence variants of genes corresponding to CyCOG 60000046 were found in different evolved strains of MED4 *Prochlorococcus*, four of which conferred resistance to cyanophages P-GSP1 and P-SSP7, whereas the other two conferred resistance only to cyanophage P-GSP1 [128]. CyCOG 60000046 encodes the enzyme UDP-glucuronate 4-epimerase, which catalyzes the inversion of the glycosyl donor isomers UDP-glucuronate and UDP-galacturonate [151]. Altering the stereochemistry of this glycan residue would affect

the molecular structure of the LPS, and presumably the binding affinity of viruses that target the LPS for adsorption. Two copies of CyCOG 60000046 were found in all isolate genomes of marine *Prochlorococcus* and *Synechococcus* (Fig 3.5B), and one copy of CyCOG 60000046 previously showed elevated expression in cultures of MED4 *Prochlorococcus* exposed to virus infection (Table 3.3). We therefore find a pattern of elevated expression of CyCOG 60000046 associated with elevated virus concentrations in both laboratory (Table 3.3) and field settings (Fig 3.1), perhaps as a result of cells switching between enzyme variants in an effort to modify LPS structure and evade infection. Alternatively, cells may simply be up-regulating LPS biosynthesis, perhaps as a means of masking other vulnerable cell surface structures. More fundamentally, the high frequency of LPS biosynthesis enzymes encoded by members of the pro2 cluster suggests that in the field, as in the lab, LPS is a significant target for viral adsorption and infection, leading to a strong selection pressure on the genes involved in determining its structure.

In some cyanobacteria, a glycoprotein S-layer forms a geometric lattice structure that is non-covalently bound to the exterior surface of the LPS [149]. S-layers are generally homogeneous tessellations of a single giant protein subunit (40 - 200 kDa) which is not highly conserved between species and can account for up to 15% of the total protein content of the cell [147]. Laboratory studies of a motile strain of marine *Synechococcus* have confirmed the presence of an S-layer encoded by the 140 kDa protein SwmA [146] and demonstrated that it plays a critical role in cell motility and susceptibility to predation by protist grazers [152, 153]. SwmA is included in CyCOG 60001830, which is the most highly-expressed CyCOG in the pro2 cluster, found in multiple copies in a subset of *Prochlorococcus* and *Synechococcus* genomes (Fig 3.5, A and B). This finding suggests that a significant subpopulation of marine *Prochlorococcus* and *Synechococcus* cells may express an S-layer in open ocean environments, and may in fact have the capacity to synthesize more than one variety of S-layer structures. S-layer proteins have been shown to be targets for phage attachment in diverse species of bacteria [154, 155], and *Clostridioides difficile* mutants expressing different S-layer protein variants exhibited resistance to different strains of phage [155], though to date no published studies have confirmed the S-layer as a phage attachment target in marine cyanobacteria. Additionally, giant proteins other than the S-layer monomer are often associated with the cell surface in prokaryotes [156], including a 1.12-MDa protein that is required for motility in a strain of marine *Synechococcus* [157], and thus giant proteins more generally are potential targets for phage attachment. Two additional CyCOGs of giant proteins ($>$ 1000 amino acids) were found in the pro2 cluster, CyCOGs 60001883 and 60001888, the former being the fourth most highly expressed CyCOG in pro2 (Fig 3.3). Genes from all three giant protein CyCOGs are situated predominantly within genomic islands and exhibit a high degree of sequence heterogeneity that precluded the generation of informative phylogenetic trees, suggesting they may be under diversifying selection pressure. Altogether, these findings imply that the S-layer and other giant proteins

may be targets for virus adsorption in marine cyanobacteria, and conversely, point to a potential role for giant proteins in a yet-to-be characterized mechanism of antiphage resistance.

Beyond LPS biosynthesis enzymes and giant proteins, there are 14 additional pro2 cluster CyCOGs that are related to cell surface structures and are predominantly transporter proteins. Exporters such as CyCOG 60001471 and proteins with autotransporter domains such as CyCOG 60003264 are putatively involved in translocating polysaccharides and peptides to the outer membrane [158, 159]. There is experimental evidence linking mutations in two transporter CyCOGs to viral resistance (Table 3.1), and similar to the LPS biosynthesis enzymes and S-layer proteins discussed, it is likely that these transporters help determine the structure of the cell surface, in this case by regulating the composition of macromolecules translocated there. Other transporters involved in essential metabolic pathways may also themselves be targets of opportunistic viruses. For instance, the "Ferrojan Horse" hypothesis predicts that some phage may take advantage of siderophore-mediated iron transporters to infiltrate and infect iron-limited marine bacteria [160]. In support of this hypothesis, we found that the substrate binding component of a putative iron transporter (CyCOG 60000706; Fig 3.3) is expressed as part of the pro2 cluster. *Prochlorococcus* and *Synechococcus* populations in the North Pacific have been shown to possess multiple iron uptake mechanisms [94], and the elevated expression of this iron transporter may be compensation for down-regulating an alternate iron transporter vulnerable to phage attachment, such as the TonB dependent siderophore receptor that is located in the outer membrane.

In addition to cell surface CyCOGs implicated in passive antiphage resistance, the pro2 cluster also included eight CyCOGs with putative roles in active antiphage defense systems. Consistent with a recent survey of bacterial antiphage defense systems in marine metagenomes [129], restriction-modification systems were the most commonly detected systems in our *Prochlorococcus* and *Synechococcus* reference genomes, and three of the eight pro2 cluster CyCOGs corresponded to restriction-modification systems (Table 3.2). The remaining five defense CyCOGs in the pro2 cluster corresponded to different types of systems that fall into three mechanistic categories: abortive infection (Lamassu, Pycsar) self/non-self discrimination systems (BREX), and systems for which the mechanism is not yet understood (Eleos, Mokosh). All of the pro2 CyCOGs linked to defense systems were part of the flexible pangenome, detected in a subset of isolate genomes (Fig 3.5). The fact that no core antiphage defense systems were found in all isolate genomes could reflect a high fitness cost associated with maintaining antiphage defense systems within a streamlined genome [23]. Alternatively, there could be core antiphage defense systems specific to marine cyanobacteria that remain to be discovered. Indeed, the CyCOG annotated as a potential component of a Lamassu defense system (CyCOG 60001376) encompasses a large family of under-characterized proteins which, though not core, are found in multiple copies across a large fraction of *Prochlorococcus* and *Synechococcus* reference genomes

(Fig 3.5C). Only 9 of the 1002 member proteins in this CyCOG were annotated as putative Lamassu system components, suggesting relatively distant homology, which could point to a novel antiphage defense system yet to be characterized. Certainly, given the high levels of phage predation faced by marine cyanobacteria, further exploration is warranted in search of novel antiphage defense systems in marine cyanobacterial genomes.

The majority of pro2 cluster CyCOGs were located within genomic islands (Fig 3.5C). This observation is consistent with previous evidence linking genomic islands to antiphage resistance and defense in *Prochlorococcus* [135, 137, 126]. Genomic islands are understood to be regions subject to increased rates of horizontal gene transfer [135, 136], and accordingly we found that the phylogenies of amino acid sequences of two pro2 CyCOGs positioned primarily within genomic islands disagreed with the taxonomic phylogeny of their genomes of origin (Fig 3.6, C and D). In contrast the phylogenies were largely concordant for two CyCOGs located outside genomic islands (Fig 3.6, A and B). The majority of the pro2 CyCOGs were also found to be part of the flexible pangenome, distributed among a subset of the cyanobacterial population, however 10 out of 100 of the pro2 appear to represent core genes, with orthologs detected in all isolate genomes in the dataset (Fig 3.5B). Notably, five of these core CyCOGs are predominantly located within genomic islands, two of which are the CyCOGs with phylogenetic signatures consistent with a high degree of horizontal gene transfer (Fig 3.6, C and D). This observation complements a previous report of a small set of core genes, primarily linked to sugar metabolism, that recruited an anomalously low number of reads from marine metagenomes, indicating a high degree of sequence variation [161]. This pattern represents a surprising departure from the traditional model of *Prochlorococcus* population diversity, in which core genes are vertically inherited, whereas horizontal gene transfer facilitates the dispersal of flexible genes [23]. Instead, there appears to be a special class of core genes that are universally detected in *Prochlorococcus* isolates, and yet highly diverse and regularly shuffled between subpopulations via horizontal gene transfer.

For both core and flexible genes, horizontal gene transfer poses a challenge for taxonomic analysis of metaomic sequence data. The phylogenetic trees of CyCOGs 60000028 and 60000046 show closely related sequence variants of the same gene detected in different clades (Fig 3.6, C and D). Other studies have found similar evidence of horizontal gene transfer across broader taxonomic divides, including nearly identical sequences detected in bacterial species pairs from different phyla [162]. Current methods assign taxonomy to sequences from mixed community environmental samples based on homology to taxonomically resolved reference genomes. This works well for vertically inherited genes for which the gene family phylogeny and the organism phylogeny are well correlated. But for horizontally transferred genes, there is no guarantee that the taxonomy of the reference sequence corresponds to the taxonomy of the organism in the sample. This may explain the increased variance we observed for flexible genes located in genomic islands when normal-

izing transcript abundance by clade as opposed to normalizing at the level of the collective cyanobacterial community (Fig 3.7, A and C). No anomalous shift in variance was apparent when comparing genus level normalization to community level normalization (Fig 3.7, B and D), indicating that the horizontal gene transfer in this cyanobacterial population may mostly occur between different clades of the same genus and not between different genera. While normalization at the community level was sufficient for the goals of study, future work may employ alternative methods such as single cell sequencing to resolve in situ taxonomies of horizontally transferred sequences.

The phenomenon of core genes that exhibit a high degree of horizontal gene transfer, which we refer to here as an "exchanged core," represents an intriguing pattern of genetic diversity that may reflect the interaction of selection dynamics with physiological constraints. Here the functional role of CyCOG 60000046 is informative, as it is an enzyme that helps determine the structure of the LPS surface of the cell wall. Experimental and observational evidence suggests that a prokaryotic strain's unique pattern of virus susceptibility is primarily determined by the structure and composition of the exterior surface of the cell wall, as this is the first point of contact between the virus and the host [137, 126, 127]. However, the cell wall is also involved in many essential cell functions beyond phage resistance, including, for example, nutrient acquisition, waste removal, and osmoregulation. Therefore, cell wall modifications that enable a host to evade viral infection have the potential to disrupt other core functions, as evidenced by experiments by Avrani et al. (2011) which demonstrated that mutations in *Prochlorococcus* genes that conferred resistance to a particular strain of phage also incurred incidental fitness costs, including slower growth and increased susceptibility to other phages [126]. Many of the genes implicated in these experiments are involved in cell wall biosynthesis, and some are members of the exchanged core CyCOGs described here (CyCOGs 60000015 and 60000046). Thus the exchanged core may be a consequence of a biological dilemma in which a subset of genes are essential for cell fitness, and yet they also present an inevitable point of vulnerability to phage predation, resulting in a diversifying selection pressure that favors a broad pool of gene variants within the wider population. Mechanisms of horizontal gene transfer, such as tycheposons that mediate exchange within genomic islands [136], may serve to maintain the genetic diversity of the exchanged core at a population level, as well as generating unique combinations of exchanged core gene variants, contributing to the collective diversity of the *Prochlorococcus* federation [23, 163]. This view is supported by the fact that both of the exchanged core genes identified here were found in multiple copies in many of the reference genomes (Fig 3.5B). Furthermore, in a subsequent study Avrani et al. (2015) found that after acquiring a resistance mutation, strains subjected to continued evolution over the course of several months accumulated compensatory mutations in other genes that partially or fully ameliorated the initial fitness cost associated with resistance [128]. In this light, some of the flexible gene content of a cell may be explained as a kind of support team that fine

tunes cell fitness in the context of the particular exchanged core variants that the cell carries. As described here, the observation of an exchanged core is consistent with the Constant Diversity theory that attributes prokaryotic population diversity to constant selection pressure from phage predation [137], and contributes new detail about the physiological drivers of the mosaic of overlapping genes and phenotypes observed in many natural populations of prokaryotes.

Within the cyanobacterial population as a whole, the pro2 cluster expression profile was elevated in the NPSG, with relatively lower expression in the northern part of the transect and the decrease coinciding with a hotspot of elevated cyanophage abundance and infection (Fig 3.2). This pattern may reflect a latitudinal shift in the population level response of North Pacific cyanobacteria to potential viral infection. For many prokaryotes, there are variable fitness costs associated with different phage response mechanisms, the relative impact of which is dependent on factors such as the diversity of viruses, infection rates, and whether the response mechanisms are constitutive or inducible [164]. It is possible that the combination of relatively lower nutrient resources and lower infection pressure in the oligotrophic NPSG leads to resistance mechanisms dominated by cell surface modifications, for which the fitness cost is fixed and can be compensated by genes in the flexible pangenome [128]. Higher infection levels in the transition zone may tip the balance in favor of defense mechanisms that can more effectively defend against specific strains of phage, but may incur higher, less predictable fitness costs, such as restriction-modification, abortive infection, and adaptive immunity systems. Alternatively, host-phage specificity may be the primary driver of the observed shift in pro2 cluster expression. The response mechanisms modeled by the pro2 cluster may be specific to the cyanophage population dominant in the NPSG, and the relative decrease in pro2 expression in the transition zone may reflect a sort of succession towards another unidentified suite of response mechanisms more effective against a different population of cyanophages dominant in that region. Further research is needed to disentangle these ecological drivers, and to characterize the impact on cyanobacterial community composition and carbon and nutrient cycles. Together, these efforts will allow future researchers to more effectively integrate the impact of phage predation and antiphage defense into ecosystem models, and to predict how these dynamics may interact with shifting climate conditions.

3.4 Methods

All code and data required to reproduce analyses and figures and to evaluate the conclusions presented in this chapter can be found in an associated repository: https://github.com/blasks/cyanophage-manuscript.

3.4.1 Coexpression pattern detection

Metatranscriptomic abundance data was processed, normalized, mapped, and tensorized as described in the previous chapter [52]. Briefly, we identified orthologous gene families based on previously defined Cyanobacterial Clusters of Orthologous Groups of proteins (CyCOGs), which consist of predicted coding sequences from reference *Prochlorococcus*, *Synechococcus* and cyanophage genomes, clustered at the amino acid level [62]. Specifically, metatranscriptomic libraries were mapped to the reference gene sequences that went into generating CyCOGs, and mapped read counts were aggregated by CyCOG and normalized. The sparse tensor decomposition model was applied to tensors of normalized residual transcript abundance data using the Barnacle Python library, as described in the previous chapter. The pro2 cluster was identified from the set of output model components on the basis of the geographic pattern of the sample weights, and the enriched KEGG pathways, as evaluated in the previous chapter. While the model was fit to the full dataset of 222 metatranscriptomic samples, the pro2 cluster showed most notable expression difference in surface waters. For this reason, pro2 expression profiles were visualized using only surface samples ($< 15m$ depth).

3.4.2 Normalization of transcript abundance data

We normalized the mapped metatranscriptomic reads using multiple methods for different parts of this study. The Barnacle clusters were generated on data normalized using version 2 of the variance stabilizing transform (vst) [64] as detailed in the previous chapter [52]. Briefly, for each sample the vst models the transcript abundance of the CyCOG as a function of overall taxon abundance, approximated by the the total number of reads recruited to the the taxon pangenome in that sample. Model residuals were output as the normalization product, indicating the degree and direction to which each CyCOG's transcript reads were detected in each sample diverged from the expectation of the vst model. For more details on this methodology refer to the Methods section of the previous chapter [52].

To evaluate the expression profiles of CyCOGs across samples in the dataset, we also normalized mapped metatranscriptomic data using a procedure outlined by Salazar et al. (2019) [51]. Briefly, transcript levels were normalized to a suite of ten universal single-copy marker genes, identified on the basis of annotation with one of the following KEGG orthologies: (K06942, K01889, K01887, K01875, K01883, K01869, K01873, K01409, K03106, and K03110). Normalization was performed at three different levels of taxonomic resolution: clade, genus, and community. For each sample, the read counts were aggregated by CyCOG for each taxon at the given resolution. For example, at the clade level, all the reads mapping to HLI reference genomes were aggregated by CyCOG, whereas for normalization at the community level, each CyCOG was aggregated across all *Prochlorococcus* and *Synechococcus* reference genomes collectively. Aggregated CyCOG

read counts were then divided by the median aggregate read count of the ten CyCOGs that corresponded to the KEGG orthologs of the ten marker genes listed. Reads mapping to virus genomes were normalized using the community level aggregated marker gene values. Samples with a median of zero marker gene reads were considered below the limit of detection and were removed from further analysis.

We compared the effect of normalization at different taxonomic resolutions by examining the variance of metatranscriptomic read counts normalized using the marker gene methodology. At each level of taxonomic resolution, the variance of each CyCOG's normalized abundances was calculated across all surface samples above the limit of detection. Variance values were then matched by CyCOG between datasets and plotted, folding in metadata specific to each CyCOG, such as the number of member gene sequences in each CyCOG, core versus flexible genes, genomic island proximity, and membership in the pro2 cluster.

3.4.3 Transcript abundance profiles

Transcript abundance profiles were constructed using the read count data normalized at the whole *Prochlorococcus* and *Synechococcus* community level. For each surface sample, normalized transcript abundance values were averaged by replicate. To generate profiles reflecting taxonomic composition, average replicate read counts were broken down into taxonomic groups (HLI *Prochlorococcus*, HLII *Prochlorococcus*, and all low light *Prochlorococcus*, other high light *Prochlorococcus*, all *Synechococcus*, and virus) based on the taxonomic assignment of the reference sequences that recruited reads. Taxonomic read counts were then: 1) summed across all samples in the dataset to generate a full dataset taxonomic profile for each CyCOG (Fig 3.1A), or 2) summed across all pro2 CyCOGs in each sample to generate a taxonomic profile for pro2 expression at each sample site (Fig 3.1, E-G).

To generate transcript abundance profiles that were comparable between CyCOGs with differing mean expression levels, we utilized a z-score transformation. For each CyCOG, the abundance profile was constructed using the replicate averaged normalized transcript abundance values for each sample, which were then subjected to z-score normalization. Heatmaps were used to visually compare z-scored transcript abundance profiles to one another (Fig 3.1, B-D). To compare profiles against the locations of various geographic and biological features of the North Pacific, z-scored abundance profiles were plotted against sample latitude (Fig 3.2). Locations for the salinity and chlorophyll fronts are from Juranek et al. (2020) [77] and the boundaries of the virus hotspot are from Carlson et al. (2022) [123].

3.4.4 Cross-referenced culture data

We cross-referenced the genes in cluster pro2 with previously published data on resistance and response to phage infection in laboratory culture studies of marine cyanobacteria. For genes shown to confer resistance

to phage infection, we compiled data from three studies: Avrani et al. (2011) [126], Marston et al. (2012) [127], and Avrani et al. (2015) [128]. We compiled data from four studies that reported alterations in host gene expression in response to exposure to phage infection, including: Lindell et al. (2007) [142], Doron et al. (2016) [143], Thompson et al. (2016) [144], and Huang et al. (2021) [145]. All gene expression data was standardized to log2 fold change over control. Even though all the genomes in these studies were included in constructing the CyCOG v6 database, different gene IDs prevented a direct mapping from reported data to CyCOG designation. We therefore mapped each reported gene to its corresponding CyCOG based on sequence similarity search. We downloaded the reference genomes reported in each publication from NCBI using the online GUI. We then used blastx (NCBI blast suite version 2.14.1) to search the downloaded nucleotide gene sequences against a database constructed from the amino acid sequences of all the proteins in CyCOG v6. We then annotated the downloaded reference gene sequences with the CyCOG identity of the closest matching CyCOG protein. Finally, the compiled culture data was aggregated by CyCOG.

3.4.5 Antiphage defense genes

We identified putative antiphage defense genes using the DefenseFinder command line application [139]. Amino acid sequences for all *Prochlorococcus* and *Synechococcus* reference sequences in CyCOG v6 [62] were collated and renamed using the format "{genome id}_{gene id}" as outlined in the DefenseFinder documentation for annotating multiple genomes simultaneously. Renamed amino acid sequences were then run through DefenseFinder (version 1.3) using the most up-to-date HMMs downloaded from the DefenseFinder site on September 7, 2024. Results were processed by taking the highest scoring hit for each gene sequence, and identifying the most common gene annotation for each CyCOG. For the vast majority of CyCOGs, there were no discordant annotations among member gene sequences. A CyCOG was considered a putative defense gene if any member genes were annotated with a DefenseFinder HMM, whereas CyCOGs were considered to be part of a putative defense system if any member gene was found to be part of a complete defense system, a more stringent criterion that requires all components of a characterized defense system to be co-localized on an annotated genome.

3.4.6 Genomic island proximity

Genomic island proximity was evaluated using genomic island locations for a set of 623 *Prochlorococcus* genomes, which were estimated by Hackl et al. (2023) using a custom expectation-maximization algorithm [136]. The *Prochlorococcus* genomes used by Hackl et al. did not fully overlap with the genomes used by Berube et al. to generate CyCOG v6 [62]. We therefore mapped genes from the Hackl et al reference genomes to CyCOGs based on sequence similarity. We used blastx (NCBI blast suite version 2.14.1) to search the

Hackl et al nucleotide gene sequences against a database constructed from the amino acid sequences of all the proteins in CyCOG v6. We then annotated the Hackl et al gene sequence with the CyCOG identity of the closest matching CyCOG protein. In cases in which reference genomres were shared in both the Hackl et al and CyCOG v6 datasets, this procedure simply matched up identical sequences between the the two datasets, allowing us to map from one set of gene ids to the other.

Once CyCOG ids had been assigned to each gene in the Hackl et al dataset, we calculated the distance of each gene to the nearest genomic island. We calculated distances as the number of base pairs between the start codon of the gene and the nearest predicted genomic island boundary. Negative distances were assigned to genes in which the start codon fell within the boundaries of a genomic island, whereas positive distances were assigned to genes with a start codon positioned outside a genomic island. CyCOG proximity distributions were then evaluated by aggregating distances by CyCOG id.

3.4.7 Gene coreness analysis

CyCOG "coreness" was calculated as the percentage of isolate genomes ($> 99\%$ complete) in which at least one ortholog of the CyCOG was detected. Coreness was evaluated on two levels of taxonomic granularity: on the level of the *Prochlorococcus* pangenome as a whole, and within more granular subgroupings (HLI *Prochlorococcus*, HLII *Prochlorococcus*, and all low light *Prochlorococcus*, other high light *Prochlorococcus*, all *Synechococcus*, and virus). For analyses or visualizations in which a binary classification was useful, all CyCOGs with a coreness of 100% were designated as "core", and all other CyCOGs were designated as "flexible".

3.4.8 Phylogenetic analysis

Phylogenetic analysis was conducted using amino acid sequences from isolate genomes used in constructing CyCOG v6. Within each analyzed CyCOG group we first removed sequences more than one standard deviation away from the median sequence length, to filter out truncated or otherwise problematic gene sequences. We used the remaining sequences to construct a multiple sequence alignment with the MUSCLE alignment software (version 5.1) [165]. Alignments were then visualized and trimmed to remove large regions of gapped alignment at the beginning or end of the alignment. Following trimming, we used RaxML (version 8.2.13) [166] to generate bootstrapped maximum likelihood phylogenetic trees, with options set to 100 bootstraps and the "PROTGAMMAWAG" substitution model. To visualize trees, we used the browser program iTOL, along with custom annotation files generated based on genome metadata and a custom Python script.

3.4.9 Statistical analyses

All statistical analyses were performed using the Python "scipy.stats" module. The "fisher_exact" function was used to test for enrichment of different CyCOG categories in the pro2 cluster, with the alternative hypothesis set to 'greater'. The test compared the proportion of CyCOGs in the pro2 cluster with an annotation of interest (e.g. genes shown to confer resistance to viral infection) to the proportion of CyCOGs with that annotation in the pangenome as a whole. The odds ratio was reported as the enrichment over expectation, given random sampling of the pro2 CyCOGs from the pangenome background, and the p-value was used to designate significance, with a threshold of $p < 0.05$.

The "mannwhitneyu" function was used as a non-parametric hypothesis test to assess differences in genomic island distance distributions. We ranked the median proximity of each CyCOG in the pangenome to its nearest genomic island, with the most negative proximity (indicating a position within a genomic island) given the highest rank and the largest positive proximity (indicating a position far from a genomic island) given the lowest rank. We then generated two vectors of rank values, one encompassing all CyCOGs in the pro2 cluster and the other all CyCOGs not in the cluster, and ran the test with the alternative hypothesis set to 'less'. We used the resulting p-value to designate significance, with a threshold of $p < 0.05$.

4 Significance

The accelerating consequences of anthropogenic climate change add urgency to the goal of elucidating the role of marine microbial communities in Earth's intertwined environmental systems. As discussed in the introduction, collectively phytoplankton account for an estimated 50% of global carbon fixation, and correspondingly, 50% of global oxygen production [19]. The conditions under which this massive churn of primary productivity takes place are changing. Among the many oceanic effects of increased atmospheric carbon dioxide concentrations, seawater has already acidified by about 0.1 pH units globally [167], global sea ice is declining [168], and oxygen minimum zones have expanded over the past 50 years[169]. And yet, despite the central climate regulating role of the biological carbon pump, it remains unclear how global phytoplankton populations will ultimately be affected by these changing conditions. By one estimate that relies on in situ chlorophyll measurements, global phytoplankton numbers have already declined over the past century, potentially as a result of rising sea surface temperatures [170]. Another study that analyzed satellite image data concluded the opposite: that global chlorophyll levels have actually increased, mainly as a result of changing seasonal patterns at the tropics [171]. Similarly, different models predict opposite trends for the global population of *Prochlorococcus* under similar future ocean projections [172]. For the sake of accurate climate models, resolving these conflicting accounts is a scientific imperative; one that hinges on a deeper understanding of microbial physiology and ecology.

In addition to informing climate forecasts, improved models of marine microbial ecology are necessary to inform growing public interest in ocean-based carbon dioxide removal technologies. For instance, ocean alkalinity engineering has attracted renewed interest in recent years because of its conceptual simplicity and potential to sequester mass quantities of carbon dioxide. The approach proposes to supplement natural inputs of carbonate and silicate minerals to the surface ocean, which cause dissolved carbon dioxide to be locked into carbonate or bicarbonate ions, ultimately shifting the sea-air equilibrium towards greater ocean absorption of inorganic carbon [173]. Initial small scale studies suggest that key groups of phytoplankton may be resilient to alkalinity enhancement [174], however more research is needed to more broadly characterize potential ecological consequences and local impacts at input sites. One of the more controversial carbon dioxide removal technologies, known as ocean fertilization, proposes to increase carbon sequestration by stimulating phytoplankton blooms with the input of iron or other limiting nutrients [175]. While some fertilization experiments suggest that under particular conditions a portion of the carbon biologically fixed during blooms may be sequestered for sufficiently long time periods [176], there is growing consensus that the relative efficacy compared to other negative emissions technologies is modest at best, and offset by significant risks of ecosystem destabilization [175]. Other notable technologies include artificial upwelling

[177], seaweed cultivation [178], and direct carbon dioxide removal from seawater [179], each presenting a unique set of considerations. Of course, any intervention in natural systems comes with serious ecological risks, to say nothing of the socioeconomic risks connected to fisheries and other ocean services. Nevertheless, there is a general consensus among climate experts that in addition to reducing carbon emissions to zero, it will be necessary to actively remove carbon dioxide from the atmosphere in order to mitigate the worst effects of climate change [180]. As climate impacts worsen and public pressure grows to implement large scale carbon removal technologies, it is critical that ecologists and oceanographers are ready to inform decision makers of the potential ecological and biogeochemical consequences of various approaches.

The molecular biology of marine microbiomes is also increasingly recognized as a resource for developing technologies to mitigate the harms of anthropogenic climate change, and to promote human flourishing more broadly. This year marks the 40th issue of an annual review of marine natural products, which routinely reports the discovery and characterization of thousands of new compounds from marine organisms and ecosystems [181]. A large portion of these natural products are produced by phytoplankton and other marine microbes, and some have been developed as ingredients in drugs [182], food technologies [183], and cosmetics [184]. Additionally, while the economic success of algal-derived biofuels has yet to be realized at scale, phytoplankton continue to show great potential for the production of biodiesel, in part because of their greater oil content, compared to terrestrial plant crops [185]. Similarly, marine microbes are of interest as potential platforms for smaller scale bioproduction of high-value chemical products. For example, the PCC 7002 strain of *Synechococcus* has long been studied as an potential platform organism for biosynthesis, in part because of its tolerance of a wide range of salinity and light conditions, and because as a photosynthetic organism, it requires fewer inputs in terms of feed stocks [186]. These examples are doubtless only the beginning of biotechnological innovation using marine microbiology. As we continue to explore the functional dark matter of marine microbial genetics, it is likely that newly characterized natural systems will in some cases give rise to new technologies, in the same way that the characterization of polymerases and CRISPR-Cas systems led to the development of PCR and gene editing.

Finally, the scientific and technological tools to address climate change are insufficient without the political and social power that comes from the organization and advocacy of individual citizens and their communities. I believe that the power of science to inform and inspire is critical to this movement. In particular, I think that as biologists, and indeed all scientists, we carry the responsibility to remember and to remind our communities of our absolute dependence on the ecosystems in which we live, work, and play. In her book "Braiding Sweetgrass" the Potowatomi (First Nation) botanist and author Robin Wall Kimmerer writes of an ancestral founder that "it was through her actions of reciprocity, the give and take with the land, that the original immigrant became indigenous. For all of us, becoming indigenous to a place means living

as if your children's future mattered, to take care of the land as if our lives, both material and spiritual, depended on it" [187]. Kimmerer points out that we are each actors in an ecosystem, and that as humans we have the solemn responsibility of awareness and agency over our own niche. I believe that wielding this power responsibly depends on cultivating a sense of curiosity and awe. In my doctoral research it has been my great privilege to become intimately acquainted with the inner rhythms of some of our planet's smallest and most abundant organisms. The wonder this has inspired in me is powerful, and an experience I know I share with many of my brilliant and motivated peers. My hope for myself and my fellow scientists is that we take seriously the work of fostering this sense of wonder – in ourselves, in our colleagues, and in our neighbors – so that our species may eventually find our way into a more harmonious relationship with the Earth, and her many inhabitants that so captivate our curiosity.

References

[1] H. S. Bernhardt, W. P. Tate, Primordial soup or vinaigrette: did the rna world evolve at acidic ph?, Biology direct 7 (1) (2012) 1–12.

[2] W. Gilbert, Origin of life: The rna world, nature 319 (6055) (1986) 618–618.

[3] E. J. Javaux, Challenges in evidencing the earliest traces of life, Nature 572 (7770) (2019) 451–460.

[4] W. W. Fischer, J. Hemp, J. E. Johnson, Evolution of oxygenic photosynthesis, Annual Review of Earth and Planetary Sciences 44 (2016) 647–683.

[5] B. E. Schirrmeister, M. Gugger, P. C. Donoghue, Cyanobacteria and the great oxidation event: evidence from genes and fossils, Palaeontology 58 (5) (2015) 769–785.

[6] P. G. Falkowski, M. E. Katz, A. H. Knoll, A. Quigg, J. A. Raven, O. Schofield, F. Taylor, The evolution of modern eukaryotic phytoplankton, science 305 (5682) (2004) 354–360.

[7] A. Z. Worden, M. J. Follows, S. J. Giovannoni, S. Wilken, A. E. Zimmerman, P. J. Keeling, Rethinking the marine carbon cycle: factoring in the multifarious lifestyles of microbes, Science 347 (6223) (2015) 1257594.

[8] H. Gest, The discovery of microorganisms by robert hooke and antoni van leeuwenhoek, fellows of the royal society, Notes and records of the Royal Society of London 58 (2) (2004) 187–201.

[9] J. R. Dolan, Unmasking "the eldest son of the father of protozoology": Charles king, Protist 170 (4) (2019) 374–384.

[10] S. W. Chisholm, R. J. Olson, E. R. Zettler, R. Goericke, J. B. Waterbury, N. A. Welschmeyer, A novel free-living prochlorophyte abundant in the oceanic euphotic zone, Nature 334 (6180) (1988) 340–343.

[11] G. Rocap, F. W. Larimer, J. Lamerdin, S. Malfatti, P. Chain, N. A. Ahlgren, A. Arellano, M. Coleman, L. Hauser, W. R. Hess, et al., Genome divergence in two prochlorococcus ecotypes reflects oceanic niche differentiation, Nature 424 (6952) (2003) 1042–1047.

[12] E. V. Armbrust, J. A. Berges, C. Bowler, B. R. Green, D. Martinez, N. H. Putnam, S. Zhou, A. E. Allen, K. E. Apt, M. Bechner, et al., The genome of the diatom thalassiosira pseudonana: ecology, evolution, and metabolism, Science 306 (5693) (2004) 79–86.

[13] S. J. Giovannoni, H. J. Tripp, S. Givan, M. Podar, K. L. Vergin, D. Baptista, L. Bibbs, J. Eads, T. H. Richardson, M. Noordewier, et al., Genome streamlining in a cosmopolitan oceanic bacterium, science 309 (5738) (2005) 1242–1245.

[14] Y. Zhang, K. N. Thompson, C. Huttenhower, E. A. Franzosa, Statistical approaches for differential expression analysis in metatranscriptomics, Bioinformatics 37 (Supplement_1) (2021) i34–i41.

[15] M. Shakya, C.-C. Lo, P. S. Chain, Advances and challenges in metatranscriptomic analysis, Frontiers in genetics (2019) 904.

[16] T. Prakash, T. D. Taylor, Functional assignment of metagenomic data: challenges and applications, Briefings in bioinformatics 13 (6) (2012) 711–727.

[17] R. D. Groussman, S. N. Coesel, B. P. Durham, E. V. Armbrust, Diel-regulated transcriptional cascades of microbial eukaryotes in the north pacific subtropical gyre, Frontiers in microbiology 12 (2021).

[18] G. A. Pavlopoulos, F. A. Baltoumas, S. Liu, O. Selvitopi, A. P. Camargo, S. Nayfach, A. Azad, S. Roux, L. Call, N. N. Ivanova, et al., Unraveling the functional dark matter through global metagenomics, Nature 622 (7983) (2023) 594–602.

[19] C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, Primary production of the biosphere: integrating terrestrial and oceanic components, science 281 (5374) (1998) 237–240.

[20] C. W. Cady, R. H. Crabtree, G. W. Brudvig, Functional models for the oxygen-evolving complex of photosystem ii, Coordination chemistry reviews 252 (3-4) (2008) 444–455.

[21] P. Greife, M. Schönborn, M. Capone, R. Assunção, D. Narzi, L. Guidoni, H. Dau, The electron–proton bottleneck of photosynthetic oxygen evolution, Nature 617 (7961) (2023) 623–628.

[22] C. MacGregor-Chatwin, P. Jackson, M. Sener, J. Chidgey, A. Hitchcock, P. Qian, G. Mayneord, M. Johnson, Z. Luthey-Schulten, M. Dickman, et al., Membrane organization of photosystem i complexes in the most abundant phototroph on earth, Nature plants 5 (8) (2019) 879–889.

[23] S. J. Biller, P. M. Berube, D. Lindell, S. W. Chisholm, Prochlorococcus: the structure and function of collective diversity, Nature Reviews Microbiology 13 (1) (2015) 13–27.

[24] C. J. Conselice, A. Wilkinson, K. Duncan, A. Mortlock, The evolution of galaxy number density at z¡ 8 and its implications, The Astrophysical Journal 830 (2) (2016) 83.

[25] N. Prywes, N. R. Phillips, O. T. Tuck, L. E. Valentin-Alvarado, D. F. Savage, Rubisco function, evolution, and engineering, Annual review of biochemistry 92 (1) (2023) 385–410.

[26] Y. M. Bar-On, R. Phillips, R. Milo, The biomass distribution on earth, Proceedings of the National Academy of Sciences 115 (25) (2018) 6506–6511.

[27] C. A. Suttle, Marine viruses—major players in the global ecosystem, Nature reviews microbiology 5 (10) (2007) 801–812.

[28] D. A. Hansell, Recalcitrant dissolved organic carbon fractions, Annual review of marine science 5 (1) (2013) 421–445.

[29] M. A. Moran, E. B. Kujawinski, W. F. Schroer, S. A. Amin, N. R. Bates, E. M. Bertrand, R. Braakman, C. T. Brown, M. W. Covert, S. C. Doney, et al., Microbial metabolites in the marine carbon cycle, Nature Microbiology 7 (4) (2022) 508–523.

[30] N. W. Sokol, E. Slessarev, G. L. Marschmann, A. Nicolas, S. J. Blazewicz, E. L. Brodie, M. K. Firestone, M. M. Foley, R. Hestrin, B. A. Hungate, et al., Life and death in the soil microbiome: how ecological processes influence biogeochemistry, Nature Reviews Microbiology 20 (7) (2022) 415–430.

[31] D. A. Hutchins, D. G. Capone, The marine nitrogen cycle: new developments and global change, Nature Reviews Microbiology 20 (7) (2022) 401–414.

[32] D. Hutchins, P. Boyd, Marine phytoplankton and the changing ocean iron cycle, Nature Climate Change 6 (12) (2016) 1072–1079.

[33] S. Dutkiewicz, M. J. Follows, J. G. Bragg, Modeling the coupling of ocean ecology and biogeochemistry, Global Biogeochemical Cycles 23 (4) (2009).

[34] K. Fennel, J. P. Mattern, S. C. Doney, L. Bopp, A. M. Moore, B. Wang, L. Yu, Ocean biogeochemical modelling, Nature Reviews Methods Primers 2 (1) (2022) 76.

[35] M. A. Moran, The global ocean microbiome, Science 350 (6266) (2015) aac8455.

[36] Y. Ni, J. Li, G. Panagiotou, Coman: a web server for comprehensive metatranscriptomics analysis, BMC genomics 17 (1) (2016) 1–7.

[37] A. Ma, M. Sun, A. McDermaid, B. Liu, Q. Ma, Metaqubic: a computational pipeline for gene-level functional profiling of metagenome and metatranscriptome, Bioinformatics 35 (21) (2019) 4474–4477.

[38] Z. Liu, Q. Wang, A. Ma, S. Feng, D. Chung, J. Zhao, Q. Ma, B. Liu, Inference of disease-associated microbial gene modules based on metagenomic and metatranscriptomic data, Computers in Biology and Medicine 165 (2023) 107458.

[39] B. Kolody, J. McCrow, L. Z. Allen, F. Aylward, K. Fontanez, A. Moustafa, M. Moniruzzaman, F. Chavez, C. Scholin, E. Allen, et al., Diel transcriptional response of a california current plankton microbiome to light, low iron, and enduring viral infection, The ISME journal 13 (11) (2019) 2817–2833.

[40] T. G. Kolda, B. W. Bader, Tensor decompositions and applications, SIAM review 51 (3) (2009) 455–500.

[41] W. Saelens, R. Cannoodt, Y. Saeys, A comprehensive evaluation of module detection methods for gene expression data, Nature communications 9 (1) (2018) 1–12.

[42] V. Hore, A. Viñuela, A. Buil, J. Knight, M. I. McCarthy, K. Small, J. Marchini, Tensor decomposition for multiple-tissue gene expression experiments, Nature genetics 48 (9) (2016) 1094–1100.

[43] M. Wang, J. Fischer, Y. S. Song, Three-way clustering of multi-tissue multi-individual gene expression data using semi-nonnegative tensor decomposition, The annals of applied statistics 13 (2) (2019) 1103.

[44] R. Bro, Parafac. tutorial and applications, Chemometrics and intelligent laboratory systems 38 (2) (1997) 149–171.

[45] M. Roald, C. Schenker, V. D. Calhoun, T. Adali, R. Bro, J. E. Cohen, E. Acar, An AO-ADMM Approach to Constraining PARAFAC2 on All Modes, SIAM Journal on Mathematics of Data Science 4 (3) (2022) 1191–1222.

[46] J. B. Kruskal, Three-way arrays: rank and uniqueness of trilinear decompositions, with application to arithmetic complexity and statistics, Linear algebra and its applications 18 (2) (1977) 95–138.

[47] A. Beck, M. Teboulle, A fast iterative shrinkage-thresholding algorithm with application to wavelet-based image deblurring, in: 2009 IEEE International Conference on Acoustics, Speech and Signal Processing, IEEE, 2009, pp. 693–696.

[48] B. O'donoghue, E. Candes, Adaptive restart for accelerated gradient schemes, Foundations of computational mathematics 15 (2015) 715–732.

[49] T. Hastie, R. Tibshirani, J. H. Friedman, J. H. Friedman, The elements of statistical learning: data mining, inference, and prediction, Vol. 2, Springer, 2009.

[50] J. Tamames, M. Cobo-Simón, F. Puente-Sánchez, Assessing the performance of different approaches for functional and taxonomic annotation of metagenomes, BMC genomics 20 (1) (2019) 1–16.

[51] G. Salazar, L. Paoli, A. Alberti, J. Huerta-Cepas, H.-J. Ruscheweyh, M. Cuenca, C. M. Field, L. P. Coelho, C. Cruaud, S. Engelen, et al., Gene expression changes and community turnover differentially shape the global ocean metatranscriptome, Cell 179 (5) (2019) 1068–1083.

[52] S. Blaskowski, M. Roald, P. M. Berube, R. Braakman, E. V. Armbrust, Simultaneous acclimation to nitrogen and iron scarcity in open ocean cyanobacteria revealed by sparse tensor decomposition of metatranscriptomes, bioRxiv (2024) 2024–07.

[53] J. Kossaifi, Y. Panagakis, A. Anandkumar, M. Pantic, TensorLy: Tensor Learning in Python, Journal of Machine Learning Research 20 (26) (2019) 1–6.

[54] R. Cavicchioli, W. J. Ripple, K. N. Timmis, F. Azam, L. R. Bakken, M. Baylis, M. J. Behrenfeld, A. Boetius, P. W. Boyd, A. T. Classen, et al., Scientists' warning to humanity: microorganisms and climate change, Nature Reviews Microbiology 17 (9) (2019) 569–586.

[55] P. Flombaum, J. L. Gallegos, R. A. Gordillo, J. Rincón, L. L. Zabala, N. Jiao, D. M. Karl, W. K. Li, M. W. Lomas, D. Veneziano, et al., Present and future global distributions of the marine cyanobacteria prochlorococcus and synechococcus, Proceedings of the National Academy of Sciences 110 (24) (2013) 9824–9829.

[56] R. T. Letscher, J. K. Moore, A. C. Martiny, M. W. Lomas, Biodiversity and stoichiometric plasticity increase pico-phytoplankton contributions to marine net primary productivity and the biological pump, Global Biogeochemical Cycles 37 (8) (2023) e2023GB007756.

[57] D. J. Scanlan, M. Ostrowski, S. Mazard, A. Dufresne, L. Garczarek, W. R. Hess, A. F. Post, M. Hagemann, I. Paulsen, F. Partensky, Ecological genomics of marine picocyanobacteria, Microbiology and Molecular Biology Reviews 73 (2) (2009) 249–299.

[58] J. O. McInerney, A. McNally, M. J. O'connell, Why prokaryotes have pangenomes, Nature microbiology 2 (4) (2017) 1–5.

[59] A. Dufresne, M. Ostrowski, D. J. Scanlan, L. Garczarek, S. Mazard, B. P. Palenik, I. T. Paulsen, N. T. de Marsac, P. Wincker, C. Dossat, et al., Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria, Genome biology 9 (5) (2008) 1–16.

[60] T. J. Browning, C. M. Moore, Global analysis of ocean phytoplankton nutrient limitation reveals high prevalence of co-limitation, Nature Communications 14 (1) (2023) 5014.

[61] J. J. Polovina, E. A. Howell, D. R. Kobayashi, M. P. Seki, The transition zone chlorophyll front updated: advances from a decade of research, Progress in Oceanography 150 (2017) 79–85.

[62] P. M. Berube, S. J. Biller, T. Hackl, S. L. Hogle, B. M. Satinsky, J. W. Becker, R. Braakman, S. B. Collins, L. Kelly, J. Berta-Thompson, et al., Single cell genomes of prochlorococcus, synechococcus, and sympatric microbes from diverse marine environments, Scientific data 5 (1) (2018) 1–11.

[63] C. Hafemeister, R. Satija, Normalization and variance stabilization of single-cell rna-seq data using regularized negative binomial regression, Genome biology 20 (1) (2019) 1–15.

[64] S. Choudhary, R. Satija, Comparison and evaluation of statistical error models for scrna-seq, Genome biology 23 (1) (2022) 27.

[65] K. Leonhardt, N. A. Straus, An iron stress operon involved in photosynthetic electron transport in the marine cyanobacterium synechococcus sp. pcc 7002, Microbiology 138 (8) (1992) 1613–1621.

[66] J. L. Roche, P. W. Boyd, R. M. L. McKay, R. J. Geider, Flavodoxin as an in situ marker for iron stress in phytoplankton, Nature 382 (6594) (1996) 802–805.

[67] T. S. Bibby, J. Nield, J. Barber, Iron deficiency induces the formation of an antenna ring around trimeric photosystem i in cyanobacteria, Nature 412 (6848) (2001) 743–745.

[68] E. Boekema, A. Hifney, A. Yakushevska, M. Piotrowski, W. Keegstra, S. Berry, K.-P. Michel, E. Pistorius, J. Kruip, A giant chlorophyll–protein complex induced by iron deficiency in cyanobacteria, Nature 412 (6848) (2001) 745–748.

[69] E. R. Zinser, D. Lindell, Z. I. Johnson, M. E. Futschik, C. Steglich, M. L. Coleman, M. A. Wright, T. Rector, R. Steen, N. McNulty, et al., Choreography of the transcriptome, photophysiology, and cell cycle of a minimal photoautotroph, prochlorococcus, PloS one 4 (4) (2009) e5135.

[70] C. Bird, M. Wyman, Nitrate/nitrite assimilation system of the marine picoplanktonic cyanobacterium synechococcus sp. strain wh 8103: effect of nitrogen source and availability on gene expression, Applied and environmental microbiology 69 (12) (2003) 7009–7018.

[71] A. C. Tolonen, J. Aach, D. Lindell, Z. I. Johnson, T. Rector, R. Steen, G. M. Church, S. W. Chisholm, Global gene expression of prochlorococcus ecotypes in response to changes in nitrogen availability, Molecular systems biology 2 (1) (2006) 53.

[72] M. Ludwig, D. A. Bryant, Acclimation of the global transcriptome of the cyanobacterium synechococcus sp. strain pcc 7002 to nutrient limitations and different nitrogen sources, Frontiers in microbiology 3 (2012) 145.

[73] T. Bibby, I. Mary, J. Nield, F. Partensky, J. Barber, Low-light-adapted prochlorococcus species possess specific antennae for each photosystem, Nature 424 (6952) (2003) 1051–1054.

[74] A. W. Thompson, K. Huang, M. A. Saito, S. W. Chisholm, Transcriptome response of high-and low-light-adapted prochlorococcus strains to changing iron availability, The ISME journal 5 (10) (2011) 1580–1594.

[75] S. Blanco-Ameijeiras, C. Cosio, C. S. Hassler, Long-term acclimation to iron limitation reveals new insights in metabolism regulation of synechococcus sp. pcc7002, Frontiers in Marine Science 4 (2017) 247.

[76] N. E. Gilbert, G. R. LeCleir, R. F. Strzepek, M. J. Ellwood, B. S. Twining, S. Roux, C. Pennacchio, P. W. Boyd, S. W. Wilhelm, Bioavailable iron titrations reveal oceanic synechococcus ecotypes optimized for different iron availabilities, ISME Communications 2 (1) (2022) 54.

[77] L. W. Juranek, A. E. White, M. Dugenne, F. Henderikx Freitas, S. Dutkiewicz, F. Ribalet, S. Ferrón, E. V. Armbrust, D. M. Karl, The importance of the phytoplankton "middle class" to ocean net community production, Global Biogeochemical Cycles 34 (12) (2020) e2020GB006702.

[78] H. Berthelot, S. Duhamel, S. L'helguen, J.-F. Maguer, S. Wang, I. Cetinić, N. Cassar, Nanosims single cell analyses reveal the contrasting nitrogen sources for small phytoplankton, The ISME Journal 13 (3) (2019) 651–662.

[79] P. M. Berube, A. Rasmussen, R. Braakman, R. Stepanauskas, S. W. Chisholm, Emergence of trait variability through the lens of nitrogen assimilation in prochlorococcus, Elife 8 (2019) e41043.

[80] H. Doré, U. Guyet, J. Leconte, G. K. Farrant, B. Alric, M. Ratin, M. Ostrowski, M. Ferrieux, L. Brillet-Guéguen, M. Hoebeke, et al., Differential global distribution of marine picocyanobacteria gene clusters reveals distinct niche-related adaptive strategies, The ISME Journal 17 (5) (2023) 720–732.

[81] Y. Harano, I. Suzuki, S.-i. Maeda, T. Kaneko, S. Tabata, T. Omata, Identification and nitrogen regulation of the cyanase gene from the cyanobacteria synechocystis sp. strain pcc 6803 and synechococcus sp. strain pcc 7942, Journal of bacteriology 179 (18) (1997) 5744–5750.

[82] S. Klähn, C. Steglich, W. R. Hess, M. Hagemann, Glucosylglycerate: a secondary compatible solute common to marine cyanobacteria from nitrogen-poor environments, Environmental microbiology 12 (1) (2010) 83–94.

[83] K.-P. Michel, E. K. Pistorius, Adaptation of the photosynthetic electron transport chain in cyanobacteria to iron deficiency: the function of idia and isia, Physiologia plantarum 120 (1) (2004) 36–50.

[84] E. A. Webb, J. W. Moffett, J. B. Waterbury, Iron stress in open-ocean cyanobacteria (synechococcus, trichodesmium, and crocosphaera spp.): identification of the idia protein, Applied and Environmental Microbiology 67 (12) (2001) 5444–5452.

[85] J. J. Marsh, H. G. Lebherz, Fructose-bisphosphate aldolases: an evolutionary history, Trends in biochemical sciences 17 (3) (1992) 110–113.

[86] A. E. Allen, A. Moustafa, A. Montsant, A. Eckert, P. G. Kroth, C. Bowler, Evolution and functional diversification of fructose bisphosphate aldolase genes in photosynthetic marine diatoms, Molecular Biology and Evolution 29 (1) (2012) 367–379.

[87] C. Schiksnis, M. Xu, M. A. Saito, M. McIlvin, D. Moran, X. Bian, S. G. John, Q. Zheng, N. Yang, F. Fu, et al., Proteomics analysis reveals differential acclimation of coastal and oceanic synechococcus to climate warming and iron limitation, Frontiers in Microbiology 15 (2024) 1323499.

[88] D. M. Sherman, L. Sherman, Effect of iron deficiency and iron restoration on ultrastructure of anacystis nidulans, Journal of bacteriology 156 (1) (1983) 393–401.

[89] S. Shinde, X. Zhang, S. P. Singapuri, I. Kalra, X. Liu, R. M. Morgan-Kiss, X. Wang, Glycogen metabolism supports photosynthesis start through the oxidative pentose phosphate pathway in cyanobacteria, Plant physiology 182 (1) (2020) 507–517.

[90] S. Lin, J. E. Cronan, Closing in on complete pathways of biotin biosynthesis, Molecular BioSystems 7 (6) (2011) 1811–1821.

[91] V. M. Boradia, M. Raje, C. I. Raje, Protein moonlighting in iron metabolism: glyceraldehyde-3-phosphate dehydrogenase (gapdh), Biochemical Society Transactions 42 (6) (2014) 1796–1801.

[92] S. Yamazaki, J. Nomata, Y. Fujita, Differential operation of dual protochlorophyllide reductases for chlorophyll biosynthesis in response to environmental oxygen levels in the cyanobacterium leptolyngbya boryana, Plant physiology 142 (3) (2006) 911–922.

[93] S. Nambi, J. E. Long, B. B. Mishra, R. Baker, K. C. Murphy, A. J. Olive, H. P. Nguyen, S. A. Shaffer, C. M. Sassetti, The oxidative stress network of mycobacterium tuberculosis reveals coordination between radical detoxification systems, Cell host & microbe 17 (6) (2015) 829–837.

[94] S. L. Hogle, T. Hackl, R. M. Bundy, J. Park, B. Satinsky, T. Hiltunen, S. Biller, P. M. Berube, S. W. Chisholm, Siderophores as an iron source for picocyanobacteria in deep chlorophyll maximum layers of the oligotrophic ocean, The ISME Journal 16 (6) (2022) 1636–1646.

[95] N. A. Straus, Iron deprivation: physiology and gene regulation, in: The molecular biology of cyanobacteria, Springer, 1994, pp. 731–750.

[96] J. M. Fraser, S. E. Tulk, J. A. Jeans, D. A. Campbell, T. S. Bibby, A. M. Cockshutt, Photophysiological and photosynthetic complex changes during iron starvation in synechocystis sp. pcc 6803 and synechococcus elongatus pcc 7942, PLoS One 8 (3) (2013) e59861.

[97] A. Latifi, M. Ruiz, C.-C. Zhang, Oxidative stress in cyanobacteria, FEMS microbiology reviews 33 (2) (2009) 258–278.

[98] J. K. Zorz, J. R. Allanach, C. D. Murphy, M. S. Roodvoets, D. A. Campbell, A. M. Cockshutt, The rubisco to photosystem ii ratio limits the maximum photosynthetic rate in picocyanobacteria, Life 5 (1) (2015) 403–417.

[99] J. M. Sobota, J. A. Imlay, Iron enzyme ribulose-5-phosphate 3-epimerase in escherichia coli is rapidly damaged by hydrogen peroxide but can be protected by manganese, Proceedings of the National Academy of Sciences 108 (13) (2011) 5402–5407.

[100] L. Marri, G. Thieulin-Pardo, R. Lebrun, R. Puppo, M. Zaffagnini, P. Trost, B. Gontero, F. Sparla, Cp12-mediated protection of calvin–benson cycle enzymes from oxidative stress, Biochimie 97 (2014) 228–237.

[101] R. Braakman, M. J. Follows, S. W. Chisholm, Metabolic evolution and the self-organization of ecosystems, Proceedings of the National Academy of Sciences 114 (15) (2017) E3091–E3100.

[102] S. Sakata, N. Mizusawa, H. Kubota-Kawai, I. Sakurai, H. Wada, Psb28 is involved in recovery of photosystem ii at high temperature in synechocystis sp. pcc 6803, Biochimica et Biophysica Acta (BBA)-Bioenergetics 1827 (1) (2013) 50–59.

[103] H. Liu, D. A. Weisz, H. B. Pakrasi, Multiple copies of the psbq protein in a cyanobacterial photosystem ii assembly intermediate complex, Photosynthesis research 126 (2015) 375–383.

[104] P. I. Calzadilla, D. Kirilovsky, Revisiting cyanobacterial state transitions, Photochemical & Photobiological Sciences 19 (5) (2020) 585–603.

[105] A. Wilson, G. Ajlani, J.-M. Verbavatz, I. Vass, C. A. Kerfeld, D. Kirilovsky, A soluble carotenoid protein involved in phycobilisome-related energy dissipation in cyanobacteria, The Plant Cell 18 (4) (2006) 992–1007.

[106] D. J. Lea-Smith, P. Bombelli, R. Vasudevan, C. J. Howe, Photosynthetic, respiratory and extracellular electron transport pathways in cyanobacteria, Biochimica et Biophysica Acta (BBA)-Bioenergetics 1857 (3) (2016) 247–255.

[107] M. A. Saito, M. R. McIlvin, D. M. Moran, T. J. Goepfert, G. R. DiTullio, A. F. Post, C. H. Lamborg, Multiple nutrient stresses at intersecting pacific ocean biomes detected by protein biomarkers, Science 345 (6201) (2014) 1173–1177.

[108] L. J. Ustick, A. A. Larkin, C. A. Garcia, N. S. Garcia, M. L. Brock, J. A. Lee, N. A. Wiseman, J. K. Moore, A. C. Martiny, Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation, Science 372 (6539) (2021) 287–291.

[109] T. J. Browning, E. P. Achterberg, I. Rapp, A. Engel, E. M. Bertrand, A. Tagliabue, C. M. Moore, Nutrient co-limitation at the boundary of an oceanic gyre, Nature 551 (7679) (2017) 242–246.

[110] A. M. Bolger, M. Lohse, B. Usadel, Trimmomatic: a flexible trimmer for illumina sequence data, Bioinformatics 30 (15) (2014) 2114–2120.

[111] R. Patro, G. Duggal, M. I. Love, R. A. Irizarry, C. Kingsford, Salmon provides fast and bias-aware quantification of transcript expression, Nature methods 14 (4) (2017) 417–419.

[112] O. Ulloa, C. Henríquez-Castillo, S. Ramírez-Flandes, A. M. Plominsky, A. A. Murillo, C. Morgan-Lang, S. J. Hallam, R. Stepanauskas, The cyanobacterium prochlorococcus has divergent light-harvesting antennae and may have evolved in a low-oxygen ocean, Proceedings of the National Academy of Sciences 118 (11) (2021) e2025638118.

[113] L. Garczarek, U. Guyet, H. Doré, G. K. Farrant, M. Hoebeke, L. Brillet-Guéguen, A. Bisch, M. Ferrieux, J. Siltanen, E. Corre, et al., Cyanorak v2. 1: a scalable information system dedicated to the visualization and expert curation of marine and brackish picocyanobacteria genomes, Nucleic acids research 49 (D1) (2021) D667–D676.

[114] M. G. Pachiadaki, J. M. Brown, J. Brown, O. Bezuidt, P. M. Berube, S. J. Biller, N. J. Poulton, M. D. Burkart, J. J. La Clair, S. W. Chisholm, et al., Charting the complexity of the marine microbiome through single-cell genomics, Cell 179 (7) (2019) 1623–1635.

[115] F. Sievers, A. Wilm, D. Dineen, T. J. Gibson, K. Karplus, W. Li, R. Lopez, H. McWilliam, M. Remmert, J. Söding, et al., Fast, scalable generation of high-quality protein multiple sequence alignments using clustal omega, Molecular systems biology 7 (1) (2011) 539.

[116] S. Kumar, K. Tamura, M. Nei, Mega3: integrated software for molecular evolutionary genetics analysis and sequence alignment, Briefings in bioinformatics 5 (2) (2004) 150–163.

[117] M. N. Price, P. S. Dehal, A. P. Arkin, Fasttree 2–approximately maximum-likelihood trees for large alignments, PloS one 5 (3) (2010) e9490.

[118] M. Roald, Y. M. Moe, Tlviz: Visualising and analysing tensor decomposition models with python, Journal of Open Source Software 7 (79) (2022) 4754.

[119] M. D. Ashkezari, N. R. Hagen, M. Denholtz, A. Neang, T. C. Burns, R. L. Morales, C. P. Lee, C. N. Hill, E. V. Armbrust, Simons collaborative marine atlas project (simons cmap): an open-source portal to share, visualize, and analyze ocean data, Limnology and Oceanography: Methods 19 (7) (2021) 488–496.

[120] M. Breitbart, C. Bonnain, K. Malki, N. A. Sawaya, Phage puppet masters of the marine microbial realm, Nature microbiology 3 (7) (2018) 754–766.

[121] L. Cai, H. Li, J. Deng, R. Zhou, Q. Zeng, Biological interactions with prochlorococcus: implications for the marine carbon cycle, Trends in Microbiology 32 (3) (2024) 280–291.

[122] C. A. Suttle, Viruses in the sea, Nature 437 (7057) (2005) 356–361.

[123] M. C. Carlson, F. Ribalet, I. Maidanik, B. P. Durham, Y. Hulata, S. Ferrón, J. Weissenbach, N. Shamir, S. Goldin, N. Baran, et al., Viruses affect picocyanobacterial abundance and biogeography in the north pacific ocean, Nature Microbiology 7 (4) (2022) 570–580.

[124] L. R. Thompson, Q. Zeng, L. Kelly, K. H. Huang, A. U. Singer, J. Stubbe, S. W. Chisholm, Phage auxiliary metabolic genes and the redirection of cyanobacterial host carbon metabolism, Proceedings of the National Academy of Sciences 108 (39) (2011) E757–E764.

[125] N. Mruwat, M. C. Carlson, S. Goldin, F. Ribalet, S. Kirzner, Y. Hulata, S. J. Beckett, D. Shitrit, J. S. Weitz, E. V. Armbrust, et al., A single-cell polony method reveals low levels of infected prochlorococcus in oligotrophic waters despite high cyanophage abundances, The ISME Journal 15 (1) (2021) 41–54.

[126] S. Avrani, O. Wurtzel, I. Sharon, R. Sorek, D. Lindell, Genomic island variability facilitates prochlorococcus–virus coexistence, Nature 474 (7353) (2011) 604–608.

[127] M. F. Marston, F. J. Pierciey Jr, A. Shepard, G. Gearin, J. Qi, C. Yandava, S. C. Schuster, M. R. Henn, J. B. Martiny, Rapid diversification of coevolving marine synechococcus and a virus, Proceedings of the National Academy of Sciences 109 (12) (2012) 4544–4549.

[128] S. Avrani, D. Lindell, Convergent evolution toward an improved growth rate and a reduced resistance range in prochlorococcus strains resistant to phage, Proceedings of the National Academy of Sciences 112 (17) (2015) E2191–E2200.

[129] A. Beavogui, A. Lacroix, N. Wiart, J. Poulain, T. O. Delmont, L. Paoli, P. Wincker, P. H. Oliveira, The defensome of complex bacterial communities, Nature Communications 15 (1) (2024) 2146.

[130] F. Cai, S. D. Axen, C. A. Kerfeld, Evidence for the widespread distribution of crispr-cas system in the phylum cyanobacteria, RNA biology 10 (5) (2013) 687–693.

[131] S. Zborowsky, D. Lindell, Resistance in marine cyanobacteria differs against specialist and generalist cyanophages, Proceedings of the National Academy of Sciences 116 (34) (2019) 16899–16908.

[132] S. Doron, S. Melamed, G. Ofir, A. Leavitt, A. Lopatina, M. Keren, G. Amitai, R. Sorek, Systematic discovery of antiphage defense systems in the microbial pangenome, Science 359 (6379) (2018) eaar4120.

[133] H. Georjon, A. Bernheim, The highly diverse antiphage defence systems of bacteria, Nature Reviews Microbiology 21 (10) (2023) 686–700.

[134] N. Kashtan, S. E. Roggensack, J. W. Berta-Thompson, M. Grinberg, R. Stepanauskas, S. W. Chisholm, Fundamental differences in diversity and genomic population structure between atlantic and pacific prochlorococcus, The ISME Journal 11 (9) (2017) 1997–2011.

[135] M. L. Coleman, M. B. Sullivan, A. C. Martiny, C. Steglich, K. Barry, E. F. DeLong, S. W. Chisholm, Genomic islands and the ecology and evolution of prochlorococcus, science 311 (5768) (2006) 1768–1770.

[136] T. Hackl, R. Laurenceau, M. J. Ankenbrand, C. Bliem, Z. Cariani, E. Thomas, K. D. Dooley, A. A. Arellano, S. L. Hogle, P. Berube, et al., Novel integrative elements and genomic plasticity in ocean ecosystems, Cell 186 (1) (2023) 47–62.

[137] F. Rodriguez-Valera, A.-B. Martin-Cuadrado, B. Rodriguez-Brito, L. Pasic, T. F. Thingstad, F. Rohwer, A. Mira, Explaining microbial population genomics through phage predation, Nature Precedings (2009) 1–1.

[138] D. S. Snyder, B. Brahamsha, P. Azadi, B. Palenik, Structure of compositionally simple lipopolysaccharide from marine synechococcus, Journal of bacteriology 191 (17) (2009) 5499–5509.

[139] F. Tesson, A. Hervé, E. Mordret, M. Touchon, C. D'humières, J. Cury, A. Bernheim, Systematic and quantitative view of the antiviral arsenal of prokaryotes, Nature communications 13 (1) (2022) 2561.

[140] J. E. Egido, A. R. Costa, C. Aparicio-Maldonado, P.-J. Haas, S. J. Brouns, Mechanisms and clinical importance of bacteriophage resistance, FEMS microbiology reviews 46 (1) (2022) fuab048.

[141] T. Goldfarb, H. Sberro, E. Weinstock, O. Cohen, S. Doron, Y. Charpak-Amikam, S. Afik, G. Ofir, R. Sorek, Brex is a novel phage resistance system widespread in microbial genomes, The EMBO journal 34 (2) (2015) 169–183.

[142] D. Lindell, J. D. Jaffe, M. L. Coleman, M. E. Futschik, I. M. Axmann, T. Rector, G. Kettler, M. B. Sullivan, R. Steen, W. R. Hess, et al., Genome-wide expression dynamics of a marine virus and host reveal features of co-evolution, Nature 449 (7158) (2007) 83–86.

[143] S. Doron, A. Fedida, M. A. Hernández-Prieto, G. Sabehi, I. Karunker, D. Stazic, R. Feingersch, C. Steglich, M. Futschik, D. Lindell, et al., Transcriptome dynamics of a broad host-range cyanophage and its hosts, The ISME journal 10 (6) (2016) 1437–1455.

[144] L. R. Thompson, Q. Zeng, S. W. Chisholm, Gene expression patterns during light and dark infection of prochlorococcus by cyanophage, PloS one 11 (10) (2016) e0165375.

[145] S. Huang, Y. Sun, S. Zhang, L. Long, Temporal transcriptomes of a marine cyanopodovirus and its synechococcus host during infection, Microbiologyopen 10 (1) (2021) e1150.

[146] J. McCarren, J. Heuser, R. Roth, N. Yamada, M. Martone, B. Brahamsha, Inactivation of swma results in the loss of an outer cell layer in a swimming synechococcus strain, Journal of Bacteriology 187 (1) (2005) 224–230.

[147] A. Hahn, E. Schleiff, The cell envelope, The cell biology of cyanobacteria (2014) 29–87.

[148] X. Xu, I. Khudyakov, C. P. Wolk, Lipopolysaccharide dependence of cyanophage sensitivity and aerobic nitrogen fixation in anabaena sp. strain pcc 7120, Journal of bacteriology 179 (9) (1997) 2884–2891.

[149] J. Šmarda, D. Šmajs, J. Komrska, V. Krzyžánek, S-layers on cell walls of cyanobacteria, Micron 33 (3) (2002) 257–277.

[150] A. Silipo, A. Molinaro, The diversity of the core oligosaccharide in lipopolysaccharides, Endotoxins: Structure, Function and Recognition (2010) 69–99.

[151] R. Muñoz, R. López, M. De Frutos, E. García, First molecular characterization of a uridine diphosphate galacturonate 4-epimerase: an enzyme required for capsular biosynthesis in streptococcus pneumoniae type 1, Molecular microbiology 31 (2) (1999) 703–713.

[152] B. Brahamsha, An abundant cell-surface polypeptide is required for swimming by the nonflagellated marine cyanobacterium synechococcus., Proceedings of the National Academy of Sciences 93 (13) (1996) 6504–6509.

[153] S. Strom, K. Bright, K. Fredrickson, B. Brahamsha, The synechococcus cell surface protein swma increases vulnerability to predation by flagellates and ciliates, Limnology and Oceanography 62 (2) (2017) 784–794.

[154] M. Ventura, M. Callegari, L. Morelli, et al., Surface layer variations affecting phage adsorption on seven lactobacillus helveticus strains, Annali di microbiologia ed enzimologia 49 (1999) 45–54.

[155] A. L. Royer, A. A. Umansky, M.-M. Allen, J. R. Garneau, M. Ospina-Bedoya, J. A. Kirk, G. Govoni, R. P. Fagan, O. Soutourina, L.-C. Fortier, Clostridioides difficile s-layer protein a (slpa) serves as a general phage receptor, Microbiology Spectrum 11 (2) (2023) e03894–22.

[156] O. Reva, B. Tümmler, Think big–giant genes in bacteria, Environmental microbiology 10 (3) (2008) 768–777.

[157] J. McCarren, B. Brahamsha, Swmb, a 1.12-megadalton protein that is required for nonflagellar swimming motility in synechococcus, Journal of bacteriology 189 (3) (2007) 1158–1162.

[158] J.-C. Kehr, E. Dittmann, Biosynthesis and function of extracellular glycans in cyanobacteria, Life 5 (1) (2015) 164–180.

[159] D. A. Russo, J. A. Zedler, Genomic insights into cyanobacterial protein translocation systems, Biological chemistry 402 (1) (2021) 39–54.

[160] C. Bonnain, M. Breitbart, K. N. Buck, The ferrojan horse hypothesis: iron-virus interactions in the ocean, Frontiers in Marine Science 3 (2016) 82.

[161] T. O. Delmont, A. M. Eren, Linking pangenomes and metagenomes: the prochlorococcus metapangenome, PeerJ 6 (2018) e4320.

[162] M. Sheinman, K. Arkhipova, P. F. Arndt, B. E. Dutilh, R. Hermsen, F. Massip, Identical sequences found in distant genomes reveal frequent horizontal transfer across the bacterial domain, Elife 10 (2021) e62719.

[163] N. Kashtan, S. E. Roggensack, S. Rodrigue, J. W. Thompson, S. J. Biller, A. Coe, H. Ding, P. Marttinen, R. R. Malmstrom, R. Stocker, et al., Single-cell genomics reveals hundreds of coexisting subpopulations in wild prochlorococcus, Science 344 (6182) (2014) 416–420.

[164] S. van Houte, A. Buckling, E. R. Westra, Evolutionary ecology of prokaryotic immune mechanisms, Microbiology and Molecular Biology Reviews 80 (3) (2016) 745–763.

[165] R. C. Edgar, Muscle: multiple sequence alignment with high accuracy and high throughput, Nucleic acids research 32 (5) (2004) 1792–1797.

[166] A. Stamatakis, Raxml version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies, Bioinformatics 30 (9) (2014) 1312–1313.

[167] K. Caldeira, M. E. Wickett, Anthropogenic carbon and ocean ph, Nature 425 (6956) (2003) 365–365.

[168] D. L. Kirchman, X. A. G. Morán, H. Ducklow, Microbial growth in the polar oceans—role of temperature and potential impact of climate change, Nature Reviews Microbiology 7 (6) (2009) 451–459.

[169] S. Schmidtko, L. Stramma, M. Visbeck, Decline in global oceanic oxygen content during the past five decades, Nature 542 (7641) (2017) 335–339.

[170] D. G. Boyce, M. R. Lewis, B. Worm, Global phytoplankton decline over the past century, Nature 466 (7306) (2010) 591–596.

[171] D. Antoine, A. Morel, H. R. Gordon, V. F. Banzon, R. H. Evans, Bridging ocean color observations of the 1980s and 2000s in search of long-term trends, Journal of Geophysical Research: Oceans 110 (C6) (2005).

[172] V. Bian, M. Cai, C. L. Follett, Understanding opposing predictions of prochlorococcus in a changing climate, Nature Communications 14 (1) (2023) 1445.

[173] P. Renforth, G. Henderson, Assessing ocean alkalinity for carbon sequestration, Reviews of Geophysics 55 (3) (2017) 636–674.

[174] J. A. Gately, S. M. Kim, B. Jin, M. A. Brzezinski, M. D. Iglesias-Rodriguez, Coccolithophores and diatoms resilient to ocean alkalinity enhancement: A glimpse of hope?, Science Advances 9 (24) (2023) eadg6066.

[175] P. Williamson, D. W. Wallace, C. S. Law, P. W. Boyd, Y. Collos, P. Croot, K. Denman, U. Riebesell, S. Takeda, C. Vivian, Ocean fertilization for geoengineering: a review of effectiveness, environmental impacts and emerging governance, Process Safety and Environmental Protection 90 (6) (2012) 475–488.

[176] V. Smetacek, C. Klaas, V. H. Strass, P. Assmy, M. Montresor, B. Cisewski, N. Savoye, A. Webb, F. d'Ovidio, J. M. Arrieta, et al., Deep carbon export from a southern ocean iron-fertilized diatom bloom, Nature 487 (7407) (2012) 313–319.

[177] A. Oschlies, M. Pahlow, A. Yool, R. J. Matear, Climate engineering by artificial ocean upwelling: Channelling the sorcerer's apprentice, Geophysical Research Letters 37 (4) (2010).

[178] I. B. Arzeno-Soltero, B. T. Saenz, C. A. Frieder, M. C. Long, J. DeAngelo, S. J. Davis, K. A. Davis, Large global variations in the carbon dioxide removal potential of seaweed farming due to biophysical constraints, Communications Earth & Environment 4 (1) (2023) 185.

[179] S. R. Cooley, S. Klinsky, D. R. Morrow, T. Satterfield, Sociotechnical considerations about ocean carbon dioxide removal, Annual Review of Marine Science 15 (1) (2023) 41–66.

[180] S. Fuss, C. D. Jones, F. Kraxner, G. P. Peters, P. Smith, M. Tavoni, D. P. van Vuuren, J. G. Canadell, R. B. Jackson, J. Milne, et al., Research priorities for negative emissions, Environmental Research Letters 11 (11) (2016) 115007.

[181] A. R. Carroll, B. R. Copp, T. Grkovic, R. A. Keyzers, M. R. Prinsep, Marine natural products, Natural Product Reports 41 (2) (2024) 162–207.

[182] T. F. Molinski, D. S. Dalisay, S. L. Lievens, J. P. Saludes, Drug development from marine natural products, Nature reviews Drug discovery 8 (1) (2009) 69–85.

[183] R. S. Rasmussen, M. T. Morrissey, Marine biotechnology for production of food ingredients, Advances in food and nutrition research 52 (2007) 237–292.

[184] A. Rotter, M. Barbier, F. Bertoni, A. M. Bones, M. L. Cancela, J. Carlsson, M. F. Carvalho, M. Cegłowska, J. Chirivella-Martorell, M. Conk Dalay, et al., The essentials of marine biotechnology, Frontiers In marine science 8 (2021) 158.

[185] M. Morales, C. Aflalo, O. Bernard, Microalgal lipids: A review of lipids potential and quantification for 95 phytoplankton species, Biomass and Bioenergy 150 (2021) 106108.

[186] T. T. Vu, E. A. Hill, L. A. Kucek, A. E. Konopka, A. S. Beliaev, J. L. Reed, Computational evaluation of synechococcus sp. pcc 7002 metabolism for chemical production, Biotechnology journal 8 (5) (2013) 619–630.

[187] R. Kimmerer, Braiding sweetgrass: Indigenous wisdom, scientific knowledge and the teachings of plants, Milkweed editions, 2013.