LINKING ENVIRONMENTAL AND MICROBIAL PROCESSES FROM COMMUNITY TO GLOBAL SCALES

Adrienne Hoarfrost

A dissertation submitted to the faculty at the University of North Carolina at Chapel Hill in partial fulfillment of the requirements for the degree of Doctor of Philosophy in the Department of Marine Sciences in the College of Arts and Sciences

Chapel Hill
2018

Approved By:

Carol Arnosti

John Bane

C. Titus Brown

Jaye Cable

Andreas Teske

ABSTRACT

Adrienne Hoarfrost: Linking Environmental and Microbial Processes from Community to Global Scales
(Under the direction of Carol Arnosti)

Life and the environment are inextricably interconnected. From the scale of a single microbe to the entire Earth system, biological and environmental processes have coevolved over billions of years into a complex system of interactions and feedbacks that together produce the geochemical and ecological conditions we observe around us. Community-scale processes result in net biogeochemical fluxes, which vary across regional and global scales in predictable patterns. At the community, regional, and global scale, this dissertation addresses a question central to our understanding of environmental microbial systems: How do microbial community interactions with their environment govern their functional and ecological role in the ecosystem, and how do environmental conditions shape the distribution and functional capacities of microbial genetic diversity? I demonstrate that microbial carbon cycling capacities in warm core ring waters originating from the Gulf Stream during an eddy intrusion event on the Mid-Atlantic Bight continental slope are distinct from those occurring in other shelf and shelf break water masses, illuminating the relationship between marine microbial communities and physical processes at the regional scale. As these eddy intrusion events likely increase in the future, these regional scale interactions have functional and biogeochemical implications in both present and future oceans. At the global scale, I build models to accurately predict genetic diversity of the key marine heterotroph SAR86 from environmental variables, identifying five previously unrecognized ecotypes within the SAR86 clade characterized by distinct environmental distributions, and resulting in the first global-resolution projections of SAR86 ecotype biogeography. From the community to the global scale, each level of inquiry demands solutions tailored to address the key challenges and opportunities unique to it, and new approaches are brought to bear at small and large scales, developing a more effective method to measure microbial activities in sediments to expand the range of environments for which microbial activity measurements are feasible, and providing a data discovery tool that harnesses the potential of publicly available sequencing datasets to scale data-driven discovery to ever more complex microbial systems.

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

LIST OF ABBREVIATIONS

API Application Programming Interface. A means of programmatic access to a set of resources

GOS Global Ocean Survey. An earlier 2004-2007 global ocean circumnavigation and ocean sequencing effort by the Venter Institute

GUI Graphical User Interface. A visual means of interacting with a digital or online product

MIxS Minimum Information about any (x) Sequence. Sequencing metadata standards developed by the Genomic Standards Consortium, defining the mandatory and suggested metadata fields to be provided for any sequencing sample being submitted to a data repository.

SRA Sequence Read Archive. The most abundant sequencing data repository for all classes of next-generation sequencing data, hosted by the National Center for Biotechnology Information.

TARA The 2009-2012 global circumnavigation expedition by the Tara Foundation to sample globally-distributed metagenomic sequencing datasets

INTRODUCTION

Microbial communities have fundamentally shaped the geochemical conditions on Earth, in the modern era and throughout every stage of Earth's history (Falkowski et al. 2008). They are a foundational part of virtually every ecosystem, from the marine water column to the deep subsurface, across terrestrial landscapes, urban environments, and even in our own bodies. Microbial community processes play out over both short and long timescales, influencing conditions from the community to the global scale. Processes at the microbial scale result in community-scale fluxes of biogeochemical products. These fluxes vary across different communities, and the biogeography of these communities cumulatively results in global-scale biogeochemical cycles. Over time, the interactions and feedbacks between microbial and environmental processes shape the Earth's climate and geochemical evolution.

At each scale of microbial action, the key scientific goals, challenges, and solutions depend on the scale of the microbial process in question. At the community scale, microbial enzymatic activities direct the flux of chemical energy between microbes and the environment. In the context of the carbon cycle, extracellular enzymatic activities initiate the remineralization of organic carbon (Arnosti 2011). The substrate specificities of extracellular enzymes, and the functional capacities of microbial communities to hydrolyze a spectrum of organic substrates, are central to the quantity and quality of carbon cycled by heterotrophic microbial communities. Measuring microbial extracellular enzymatic activities directly in environmental samples, therefore, is an important priority for community-scale investigations of microbial impacts on biogeochemical cycling. However, such measurements can prove infeasible in some sedimentary environments, where organic substrates sorb to the sediment matrix (Lutzow et al. 2006; Theng 1979). Activity measurements rely on sufficient recovery of fluorescently labeled substrate additions (Arnosti 1996; 2003; Hoppe 1983), and are often difficult or impossible in sediments. Chapter 1 presents a solution to this obstacle, in which a competitive desorption treatment addition to enzymatic activity measurements improves recovery of fluorescent substrates, improving the accuracy and feasibility of extracellular enzymatic activity measurements in sediments. This paper has been published (Hoarfrost et al. 2017), and has been reformatted and reprinted for this dissertation.

Enzymatic activity measurements can also illuminate the link between microbial communities and physical processes at the regional scale. Differences in functional capacities have been demonstrated across latitude (Arnosti et al. 2011), with depth (Hoarfrost & Arnosti 2017; Steen et al. 2012), at the sediment-water interface (Teske et al. 2011), and between ocean regions (Arnosti & Steen 2013). While the oceans are clearly characterized by different carbon cycling capacities across ocean regions, the relationship between physical oceanographic processes and their associated enzymatic activities, as well as the biogeochemical implications under present and future ocean conditions, remain poorly understood. In the North Atlantic, warm core eddies originating from the Gulf Stream travel toward the coast, and can persist to ultimately intrude on the continental shelf break of the Mid Atlantic Bight (Gawarkiewicz et al. 2012; Zhang and Gawarkiewicz 2015). As the oceans warm, Gulf Stream meanders are becoming more pronounced, and such eddy intrusions more frequent (Andres 2016; Monim 2017; Gawarkiewicz et al. 2018). Chapter 2 investigates the microbial carbon cycling capacities of microbial communities within distinct water masses along a transect of the Mid Atlantic Bight shelf and shelf break during an eddy intrusion event. Distinct rates of activity and spectra of substrates hydrolyzed within the warm core eddy intrusion demonstrates the relationship between microbial and mesoscale processes, the interconnections between biogeochemical cycling and the environment, and the implications of these interactions in a changing climate. This publication is in preparation for *Limnology and Oceanography*.

The differences in functional capacities seen at the community and regional scales, and their close interactions with environmental processes, beg the question of the manner in which genetic diversity varies at the global scale, and how these global distributions can be predicted by environmental variables. Differences in the functional biogeography of microbial communities are mirrored by biogeographical distributions of microbial community composition (Fuhrman et al. 2008; Ladau et al. 2013; Delong et al. 2006). While enzymatic activities provide a direct measurement of functional capacities, they are time-intensive measurements, and are difficult to gather at large spatial scales. Next-generation sequencing of whole microbial communities, in contrast, are readily available at global scales (Sunagawa et al. 2015; Rusch et al. 2007), as is environmental data from a variety of historical and satellite sources. In Chapter 3, these data sources are used to identify global ecotypes within the ubiquitous marine heterotroph SAR86, and to predict the distributions of these ecotypes at a global scale. This publication is in preparation for the *ISME* journal. The identification of SAR86 ecotypes is achieved by building machine learning models to predict SAR86 gene presence from environmental variables available at global resolution. Machine learning, a branch of statistical modeling that uses computing power to iteratively "learn" relationships from data without being explicitly programmed, thrives on large datasets. As sequencing datasets become increasingly available, new modeling approaches can bring new insight to the complex systems governing microbe-environment interactions.

The number of publicly available sequencing datasets have increased exponentially in recent years, and represent a largely untapped opportunity to answer data-intensive questions in environmental microbiology and biogeochemistry. While projects such as the TARA (Sunagawa et al. 2015) and Global Ocean Survey (Rusch et al. 2007) expeditions have provided invaluable global ocean sequencing datasets, they were time-intensive and expensive efforts that are not easily replicated, and the hundreds of samples produced are insufficient for many data-intensive questions. Future efforts to apply machine learning to marine microbiology, for example, could use every marine metagenome ever produced by next-generation sequencing technologies, likely numbering in the thousands rather than hundreds. However, parsing the more than 3 million total available sequencing datasets to identify the subset of datasets that match the research priorities for an individual project is a daunting task, with few existing resources to facilitate the process (Sayers 2017; Zhu et al. 2013). Streamlining the data discovery step is therefore a crucial task for environmental microbiology, and for bioinformatics in general. Chapter 4 presents MetaSeek, a data discovery tool for sequencing data, which integrates the metadata of all publicly available next-generation sequencing datasets in the Sequence Read Archive, and provides an easy-to-use, interactive online user interface as well as programmatic API to search, filter, save, download, and share integrated sequencing datasets for any users' needs. The MetaSeek publication is in submission to the *Application Notes* section of the *Bioinformatics* journal. Because *Bioinformatics* has strict length limits on its Application Notes submissions, an additional "explainer", with more detailed descriptions of key features of the MetaSeek tool, is made available in Appendix D. By lowering the barriers to data discovery, the first step of data-intensive investigation, MetaSeek provides a means to scaling data-driven discovery, both in the geographic sense, by exposing environmental sequencing datasets (and those from other research domains) at greater spatial and temporal resolution, and in the computational sense, exposing sufficient data to disentangle the vast complexity of microbial systems and answer some of the most pressing and elusive scientific questions in bioinformatics.

The four chapters of this dissertation address a central question underlying microbial-environmental relationships at community, regional and global scales: How do environmental variables govern the distribution of microbial genetic diversity, and how do the interactions between environmental and biological processes impact the functional and biogeochemical role of microbial communities in the environment? Differences in microbial carbon cycling capacities related to regional-scale physical processes illuminate the functional relationship between microbial communities and environmental processes, and its implications for biogeochemical cycling both now and in the future. Global-scale models of microbial genetic diversity are accurately predicted by environmental variables, defining global ecotypes in an important marine heterotroph, and resulting in the first global-scale projections of gene-level biogeography. New approaches are brought to bear at both small and large scales, overcoming central challenges to understanding environmental microbiological systems from single communities to global-scale explorations, expanding the range of environments for which microbial activities can be measured, and introducing new tools and resources for data-driven discovery. Together, the following chapters explore the links between environmental and microbial processes, traversing community-scale activities to global-scale phenomena.

REFERENCES

Andres M. (2016) On the recent destabilization of the Gulf Stream path downstream of Cape Hatteras. *Geophys Res Lett*; **43**.

Arnosti C. (1996) A new method for measuring polysaccharide hydrolysis rates in marine environments. *Org Geochem*; **25**:105–115.

Arnosti C. (2003) Fluorescent derivatization of polysaccharides and carbohydrate-containing biopolymers for measurement of enzyme activities in complex media. *J Chromatogr B Analyt Technol Biomed Life Sci*; **793**:181–91.

Arnosti C. (2011) Microbial extracellular enzymes and the marine carbon cycle. *Ann Rev Mar Sci*; **3**:401–425.

Arnosti C, Steen AD. (2013) Patterns of extracellular enzyme activities and microbial metabolism in an Arctic fjord of Svalbard and in the northern Gulf of Mexico: contrasts in carbon processing by pelagic microbial communities. *Front Microbiol*; **4**.

Delong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N, et al. (2006) Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science*; **311**:496–503.

Falkowski PG, Fenchel T, Delong EF. (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science*; **320**:1034–9.

Fuhrman JA, Steele J a, Hewson I, Schwalbach MS, Brown M V, Green JL, et al. (2008) A latitudinal diversity gradient in planktonic marine bacteria. *Proc Natl Acad Sci*; **105**:7774–8.

Gawarkiewicz, G., R. Todd, A. Plueddemann, and M. Andres. 2012. Direct interaction between the Gulf Stream and the shelf break south of New England. Sci. Rep. **2**: 477. doi:doi:10.1038/srep00553

Gawarkiewicz G, Todd RE, Zhang W, Partida J, Gangopadhyay A, Monim M-U-H, et al. (2018) The changing nature of shelfbreak exchange revealed by the OOI Pioneer Array. *Oceanography*; **31**:60–70.

Hoarfrost A, Snider R, Arnosti C. (2017) Improved measurement of extracellular enzymatic activities in subsurface sediments using competitive desorption treatment. *Front Earth Sci*; **5**:13. doi: 10.3389/feart.2017.00013.

Hoppe H. (1983) Significance of exoenzymatic activities in the ecology of brackish water: measurements by means of methylumbelliferyl-substrates. *Mar Ecol Prog Ser*; **11**:299–308.

Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, et al. (2013) Global marine bacterial diversity peaks at high latitudes in winter. *ISME J*; **7**:1669–77.

Lutzow, M. V., Kogel-Knabner, I., Ekschmitt, K., Matzner, E., Guggenberger, G., Marschner, B., et al. (2006) Stabilization of organic matter in temperate soils: Mechanisms and their relevance under different soil conditions - A review. *Eur. J. Soil Sci.;* **57**:426–445. doi:10.1111/j.1365-2389.2006.00809.x.

Monim M. (2017) Seasonal and Inter-annual Variability of Gulf Stream Warm Core Rings from 2000 to 2016. University of Massachusetts-Dartmouth.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. (2007) The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol*; **5**:0398–0431.

Sayers E. (2017). The E-utilities In-Depth: Parameters, Syntax and More. 2009 May 29 [Updated 2017 Nov 1]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US).

Steen AD, Ziervogel K, Ghobrial S, Arnosti C. (2012) Functional variation among polysaccharide-hydrolyzing microbial communities in the Gulf of Mexico. *Mar Chem*; **138–139**: 13–20.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. (2015) Structure and function of the global ocean microbiome. *Science*; **348**:1–10.

Teske A, Durbin a, Ziervogel K, Cox C, Arnosti C. (2011) Microbial community composition and function in permanently cold seawater and sediments from an arctic fjord of svalbard. *Appl Environ Microbiol*; **77**:2008–18.

Theng, B. K. G. (1979). Clay-Polymer Interactions: Summary and Perspectives. *Clays Clay Miner;* **30**:1–10. doi:10.1346/CCMN.1982.0300101.

Zhang WG, Gawarkiewicz G. (2015) Dynamics of the direct intrusion of Gulf Stream ring water onto the Mid-Atlantic Bight shelf. *Geophys Res Lett*; **42**:7687–7695.

Zhu Y, Stephens RM, Meltzer PS, Davis SR. (2013) SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics;* **14**:19.

CHAPTER 1: IMPROVED MEASUREMENT OF EXTRACELLULAR ENZYMATIC ACTIVITIES IN SUBSURFACE SEDIMENTS USING COMPETITIVE DESORPTION TREATMENT[1]

Adrienne Hoarfrost[1*], Rachel Snider[1], Carol Arnosti[1]

[1]University of North Carolina, Department of Marine Sciences, Chapel Hill, NC, USA

*** Correspondence:**

Adrienne Hoarfrost

adrienne.l.hoarfrost@unc.edu

Keywords: carbon degradation$_1$, microbial activity$_2$, deep biosphere$_3$, hydrolysis$_4$, polysaccharides$_5$.

1 Introduction

Heterotrophic microbial communities play an important role in organic carbon cycling in subsurface sediments. Increasing genomic evidence of the predominance of heterotrophy in the subsurface environment (Biddle et al., 2006; Fry et al., 2008; Lloyd et al., 2013) suggests that

[1] Hoarfrost A, Snider R, Arnosti C. (2017) Improved measurement of extracellular enzymatic activities in subsurface sediments using competitive desorption treatment. *Front Earth Sci*; **5**:13. doi: 10.3389/feart.2017.00013.

heterotrophic remineralization of organic matter plays a larger role in the sedimentary environment than previously appreciated. A key first step in the heterotrophic breakdown of organic carbon is extracellular enzymatic hydrolysis, in which compounds too large to be taken up directly are hydrolyzed to sizes small enough for transport into the cell. The need for measurements of enzymatic activities to quantify heterotrophic processes in subsurface sediments is evident, but the technical challenges associated with these measurements are such that comparatively few measurements have been made, particularly in subsurface environments. Extracellular enzymatic activity is typically measured by addition of a fluorescently labeled substrate to an environmental sample, and hydrolysis is detected either as an increase in fluorescence as a fluorophore is cleaved (Hoppe, 1983) or as a change in molecular weight distribution as a fluorescent substrate is hydrolyzed into lower molecular weight products (Arnosti, 1996, 2003). In both cases, adequate recovery of the amended label or labeled substrate is necessary for interpretable results.

However, adequate recovery of fluorescent labels or labeled substrates is often difficult to achieve due to the tendency of organic compounds to sorb strongly to the sediment matrix. Sorption occurs when the organic substrates interact with sediment surfaces. Interaction mechanisms can include ligand exchange, cation bridges, or weak interactions including hydrophobic interactions, van der Waals forces, or H-bonding (Lutzow et al., 2006; Theng, 1979). High molecular weight substrates often adsorb more strongly than low molecular weight compounds (Podoll et al., 1987) and so pose a particular challenge for activity measurements, yet these measurements are especially important as most natural organic matter is biosynthesized as high molecular weight compounds.

The strength of sorption is dependent on the characteristics of the organic substrate as well as the sediment composition and mineralogy (Kaiser and Guggenberger, 2000). The interaction of these factors leads to great variation in the extent to which enzyme activities in sediments can be measured: in some sediments, activities are measured relatively easily, in other sediments sorption affects the quality of measurements to an extent that may affect the accuracy of results, and in some sediments there is such poor recovery of substrate that measurements of enzyme activities are not feasible. Thus, published data likely excludes sediments for which measurements are particularly challenging to obtain. Measuring activities in such sediments may be important for capturing the range of heterotrophic activities in the subsurface, however, and exclusion of such sediments may bias our understanding of microbial effects on environmental processes.

Several strategies have been used to attempt to overcome the effect of sorption on measurements of enzyme activity in sediments. Very dilute sediment slurries, for example 20:1 ratios of aqueous media to sediment, can be used to minimize sorption surface area relative to substrate concentration (Lloyd et al., 2013). Another approach has attempted to correct for adsorption by calculating the sorption affinity constant of the target molecule from a separate set of incubation standards with known concentrations of fluorophore, and back-calculating the total concentration of substrate hydrolyzed in the enzyme activity calculation (Coolen et al., 2002; Coolen and Overmann, 2000). Both of these approaches have limitations, however. High dilution of sediments necessarily reduces microbial interactions with the sediment matrix, but such interactions may be important, since the interactions of organic matter with sediment particles can affect the bioavailability of substrates (Chenu and Stotzky, 2002; Keil et al., 1994). Moreover, phenomena such as quorum sensing are dependent upon close spatial interactions of organisms and substrates (Hmelo et al., 2011), so quorum-sensing dependent enzymatic activities likely could not be measured in a dilute slurry. Ideally, experimental conditions should reflect natural conditions as much as possible. Correcting for sorption, aside from requiring additional time and resources to conduct incubations for sorption affinity constant calculations, assumes that sorption is at equilibrium within 8 hours. Sorption can occur on both short and long timescales (Pignatello and Xing, 1996), however, and the factors affecting this vary by sediment type and characteristics, so correcting enzyme activities using sorption affinity constants may not always yield accurate results.

Here we present an alternative strategy to measure enzymatic activities in sediments, an approach that counteracts the effect of sorption by recovering adsorbed substrate. We adapted a method developed previously to measure extracellular enzymatic hydrolysis of high molecular weight organic matter in sediments and seawater (Arnosti, 1996, 2003). The original method involves addition of a fluorescently labeled, high molecular weight substrate to sediments. After incubation, sediment subsamples are centrifuged to obtain porewater containing the partially-hydrolyzed substrate, which are analyzed chromatographically to determine the molecular weight distribution of the hydrolysis products and thereby the hydrolysis rate. We have extended this method by developing a treatment to desorb amended labeled substrate from subsamples for better detection of enzymatic activities. We tested two desorption strategies, treatment of sediment slurry subsamples with extraction solutions at elevated pH, and treatment of subsamples with extraction solutions using competitive desorption. Elevated pH was tested because adsorption via ligand exchange occurs most strongly at acidic pH (Gu et al., 1994; Kaiser and Guggenberger, 2000), and compounds bound by this means may be more easily desorbed at high pH (Kaiser and Zech, 1999). Competitive desorption, addition of unlabeled substrate to a subsample in order to desorb the adsorbed fluorescently-labeled substrate, was tested since adsorption occurs when compounds compete to adsorb to a limited number of available sorption sites in the sediment matrix (Gu et al., 1994). The adsorption of a particular molecule is often reversible, and a given molecule can be displaced by other molecules that compete for the same sorption sites (Gu et al., 1994, 1996). Competitive displacement of adsorbed compounds has been demonstrated with mixtures of natural organic matter of similar or stronger adsorption affinities (Gu et al., 1996). We tested both pH and competitive desorption strategies, optimized a desorption extraction method, and demonstrated its effectiveness with a range of marine subsurface sediments. Here, we report the efficacy of our optimized extraction method and its applicability to subsurface sediments from a range of geochemical settings.

2 Materials and methods

2.1 Sediment collection and characteristics

Sediments for development of the extraction treatment protocol (see below) were collected from the Marmara Sea. Once finalized, the extraction treatment was applied to sediments from a range of geochemical environments in the Eastern Mediterranean Sea and the Guaymas Basin.

2.1.1 Marmara Sea and Eastern Mediterranean Sediments

Sediment from the Marmara Sea and the Eastern Mediterranean were collected during R/V *Meteor* cruise M84 in February 2011 (SI Table 1.1). Surficial sediments from the Marmara Sea (40°47.97’ N, 27°43.49’ E, 600m water depth) were collected by multicorer, and sediments from 570-585cm and 520-530cm depth horizons were collected by gravity corer. Sediments from the Eastern Mediterranean (33°02.00’ N, 32°38.00’ E, 1424m water depth) were collected by gravity corer at 365cm, 385cm, 440cm, 455cm, 575cm, and 582-590cm depth horizons. Individual depth intervals were subsampled from the cores into 50 mL centrifuge tubes, which were stored at 4°C in anaerobic chambers until use. Eastern Mediterranean sediments contained five sapropel layers that were cross-referenced with those described by Calvert and Fortugne (2001). Those used in these experiments included S4 (from 385cm), S5 (455cm), and S7 (582-590cm).

2.1.2 Guaymas Basin Sediments

Sediments from the Guaymas Basin, a spreading center within the Gulf of California, were collected aboard the R/V *El Puma* in October 2014 (Buckley et al., 2015). Sediments were collected at 5cm and 55cm sediment depth at six locations (P1, P3, P5, P8, P10, and P13) that vary in geological and environmental context (SI Table 1.1). Sediment intervals from cores were subsampled into airtight plastic containers and stored at 4°C until use in incubations.

2.2 Sediment incubation preparation

Incubations with Marmara Sea sediment were used for initial development of the extraction treatment protocol in three preliminary experiments – PreX1, PreX2, and PreX3 – using sediments from 0-5cm, 570-585cm, and 520-530cm depth intervals, respectively. Autoclaved artificial seawater was added to homogenized sediments to make a 2:1 seawater:sediment slurry. 21mL of slurry was dispensed into each of two 50mL-volume serum vials; one vial was autoclaved as a killed control, and one vial was used as a live experimental vial. These incubations were set up under aerobic conditions, and due to limited availability of sediments only chondroitin was used as a substrate.

In Eastern Mediterranean and Guaymas Basin sediments, all sample preparation was carried out in an anaerobic chamber under N_2 atmosphere. Each sediment sample was homogenized in a sterile beaker with a sterile spatula. Artificial seawater (Sigma S9883), autoclaved and cooled under N_2, was added in a 2:1 ratio to homogenized sediments and mixed thoroughly. 21mL of sediment slurry was portioned into each 50mL-volume, sterile serum vial using a sterile serological pipette, and sealed with a stopper and crimp. Nine serum vials were prepared from each sediment section – three live incubations and one killed control for each of two substrates (chondroitin and laminarin), and a live blank control. The sealed vials were removed from the anaerobic chamber, and two vials were autoclaved for 30 minutes, then cooled to room temperature to serve as killed controls. Substrate addition and subsequent subsampling of the incubations was carried out by opening the serum vials under a stream of $N_{2,}$ using aseptic technique. Substrate was added in 175µM monomer-equivalent concentrations; three of the live incubations and one killed control received fluorescently labeled chondroitin sulfate; three live incubations and one killed control received fluorescently labeled laminarin, and one live incubation served as a blank and did not receive substrate. Time zero samples were collected immediately after substrate addition; vials were then resealed with stoppers, crimped and stored at 4°C in the dark until further subsampling. Subsamples were taken at 3, 6, and 9 week timepoints.

2.3 Development of the extraction protocol

The three preliminary experiments – PreX1, PreX2, and PreX3 – were used to develop and optimize the extraction treatment protocol. In each case, fluorescently-labeled chondroitin substrate was added to incubations at 175µM monomer-equivalent concentrations (PreX1) or 350µM monomer-equivalent concentrations (PreX2 and PreX3), which is the concentration typically used in previous slurry incubations (e.g. Arnosti, 2003, 2008).

In the first experiment (PreX1), three desorption conditions were tested: competitive desorption, desorption with solution at pH=10, and desorption with solution at pH=11. The incubations were set up as described in section 2.2, and the fluorescently-labeled chondroitin substrate was added. Subsamples were taken at t0 and 2 days. At each subsampling point, 0.5mL of sediment slurry was removed from each vial and added to a treatment tube containing 2mL of either 700μM unlabeled chondroitin (competitive desorption), carbonate buffer at pH=10 (pH10), carbonate buffer at pH=11 (pH11), or a no-treatment control of 2mL DI H_2O. The no-treatment tubes were immediately centrifuged, and the supernatant was filtered through a 0.2 um pore-size filter and stored at -20°C. The treatment tubes were incubated and periodically shaken for two hours in a 30°C waterbath before centrifuging, filtering, and storing. Based on the results (see Results), competitive desorption was selected for use in subsequent experiments.

In the second experiment, PreX2, multiple concentrations of unlabeled substrate were tested for use in the competitive desorption approach. Three sediment incubations were again set up in serum vials – a live incubation, a killed incubation, and a live blank incubation, and fluorescently-labeled chondroitin substrate was added. Subsamples were taken at t0, 2 days, and 6 days. At each timepoint, 0.5mL of sediment slurry was added to each treatment tube containing 2mL of a solution of 700μM unlabeled chondroitin, 1400μM unlabeled chondroitin, or 2800μM unlabeled chondroitin. There was also a no-treatment DI H_2O control. No-treatment tubes were immediately processed, while treatment tubes were incubated and periodically shaken for two hours in a 30°C waterbath before processing. Based on the results, 2800μM concentrations of unlabeled chondroitin was selected for subsequent experiments.

The third experiment, PreX3, tested whether the addition of sodium dodecyl sulfate (SDS) provided additional improvement to the competitive desorption treatment method developed in PreX2 and PreX1. Three sediment incubations, a live, a kill, and a live blank, were used. Subsamples were taken at t0, 7 days, and 14 days. At each timepoint, 0.5mL of sediment slurry was added to each treatment tube containing 2mL of either 2800μM unlabeled chondroitin and 0.2% SDS, 2800μM unlabeled chondroitin only, or a no-treatment control of DI H_2O. No-treatment tubes were immediately processed, and treatment tubes were incubated for 2 hrs in a 30°C waterbath before processing. Competitive desorption with addition of SDS was selected as the final extraction treatment method, and was applied to the sediments from the Eastern Mediterranean and Guaymas Basin.

2.4 Activity measurements with competitive desorption treatment

At each timepoint, subsamples were taken from each incubation to measure the potential activity of extracellular enzymes that hydrolyze chondroitin or laminarin (Fig 1.1). For each subsample, two 15mL centrifuge tubes were prepared for an extraction treatment and for a no-treatment control, for a total of 18 falcon tubes per time point. Treatment tubes contained 0.5mL of 14mM unlabeled chondroitin or laminarin (2800μM in 2.5mL), 0.5mL 0.5% SDS (0.2% in 2.5mL), and 1mL DI H_2O. No-treatment tubes contained 2mL of DI H_2O.

1mL of sediment slurry was removed with a N_2-flushed syringe from each serum vial under a N_2 stream using aseptic technique, and 0.5mL of slurry was added to each of the treatment and no-treatment tubes. No-treatment tubes were immediately centrifuged (2000rpm, 4 minutes), and the supernatant was filtered through a 0.2μm pore size cellulose acetate syringe filter (Sterlitech CA0225) and stored in an epi tube at -20°C until analysis. Treatment tubes were allowed to process in a 30°C waterbath for two hours, shaking manually every 10-15 minutes, to allow desorption to occur. The treatment tubes were then centrifuged and syringe filtered in the same manner as the no-treatment tubes, and stored at -20°C.

The proportion of fluorescently-labeled substrate that had been hydrolyzed into lower-molecular-weight products in each subsample was analyzed using gel permeation chromatography with fluorescence detection, after Arnosti (1996; 2003).

2.5 Fluorescent substrate preparation and chromatogram interpretation

The substrates laminarin and chondroitin were labeled with the fluorophore fluorosceinamine after the method of Arnosti (1996, 2003). In short, hydroxyl groups at multiple sites along the substrate are activated with cyanogen bromide, then coupled with the fluorophore fluoresceinamine, resulting in a high molecular weight substrate labeled with a fluorescent label, typically at multiple positions. The molecular weight distribution of a fluorescently-labeled substrate can be visualized using gel permeation chromatography with fluorescence detection. When a live incubation is amended with the substrate, hydrolytic activity shifts the molecular weight distribution of the fluorescent substrate from all high- to a mixture of high- and lower-molecular-weight hydrolysis products, and the hydrolysis rate can then be calculated from the change in molecular weight distribution (relative to standards of known molecular weight).
To visualize the molecular weight distribution of the substrate and any hydrolysis products, a sample is injected onto a 21cm G50 Sephadex gel permeation chromatography column connected in series to a 19cm G75 Sephadex column. These columns separate a sample by molecular weight such that the highest molecular weight compounds are excluded from the pores within a gel and the lower molecular weight compounds penetrate through the pores of the gel. The higher molecular weight compounds thus elute first from the columns, while the lower molecular weight compounds elute later. Standards of known molecular weight are used to determine elution times for different molecular weights. Elution time per sample in this study was 75 minutes, at a flow rate of 1mL/min. Fluorescence of the column effluent was tracked at an emission wavelength of 530 nm (excitation at 490 nm) using a Hitachi fluorescence detector,

and the molecular weight distribution was determined from the final chromatogram output of fluorescence signal vs. time. Hydrolysis rates were calculated from the change in molecular weight distribution from time zero to the time of sampling.

The added substrates, chondroitin and laminarin, are polysaccharides with different structures and characteristics: laminarin, a storage glucan in brown algae and diatoms, is a branched polymer of β-linked glucose units, while chondroitin is a sulfated polymer of n-acetyl glucosamine and glucuronic acid. The enzymes required to hydrolyze laminarin and chondroitin sulfate have been identified in marine bacteria (Alderkamp et al., 2007; Wegner et al., 2013; Xing et al., 2015), and activities of enzymes hydrolyzing these polysaccharides have been measured in a wide range of environments (e.g. Arnosti, 2008; Arnosti et al., 2009).

2.6 Statistical analyses

When comparing whether the treatment resulted in increased total fluorescence intensity relative to no treatment in raw fluorescence units (FU; the detector signal in millivolts), a paired, one-sided t-test was used to compare chromatographic fluorescence intensities of incubations subjected to no treatment and treatment conditions. When comparing a percent improvement relative to zero, the no-treatment value was subtracted from the treatment value for a particular incubation, so an unpaired, one-sided t-test was used to test whether the percent improvement was greater than zero.

2.7 Reproducibility

The raw data from this project is stored in the BCO-DMO database (Hoarfrost and Arnosti, 2016). The scripts used to process and analyze the data, and generate the figures in this publication, can be found at the corresponding github repository (Hoarfrost, 2016).

3 Results

The extraction treatment presented here was developed to reduce the effects of adsorption on substrate recovery when measuring extracellular enzymatic activity in sediments using a fluorescently-labeled high molecular weight substrate, and to broaden the range of sediments in which enzyme activities can be measured using these substrates. Competitive desorption with unlabeled substrate and SDS proved to be effective in improving key chromatogram characteristics, by decreasing peak width and increasing fluorescence intensities (Fig 1.2, SI Fig 1.1). Some of the improvements in chromatogram characteristics can be summarized by the difference in area under the chromatogram, referred to here as the total integrated fluorescence intensity, between treatment and no treatment controls, which was used as an overall measure of chromatogram quality (Fig 1.2, Fig 1.3a and b, Fig 1.4, Fig 1.5).

3.1 Competitive desorption treatment effects on chromatogram quality

At all timepoints, desorption treatment improved several chromatogram characteristics (Fig 1.2, SI Fig 1.1). Overall, total integrated fluorescence intensities were higher in treatment relative to no treatment controls (Fig 1.3a), as can be seen by the difference in peak heights (Fig 1.2; note difference in scales on y axes). The desorption treatment was especially effective at desorbing the high molecular weight portion of the added substrate (Fig 1.3b), resulting in higher proportions of high- to low- molecular weight substrate, an effect that is particularly evident for laminarin in core P13 from Guaymas Basin (Fig 1.2b and c). Finally, the chromatogram peaks are sharper and peak width is narrower, as exemplified by the incubations with chondroitin in Mediterranean 385cm sediments (Fig 1.2a and c.). These characteristics result in higher quality chromatograms and lead to more easily interpretable rate calculations. In some cases, samples with no treatment applied resulted in very poor recovery of substrate and such low chromatogram intensities that they would be unusable (e.g. Fig 1.2b, panel 1-t0). In these cases, competitive desorption treatment enables measurement of enzymatic activities in sediments where such measurements otherwise could not be made.

Although the extraction protocol includes a 2-hour incubation of a subsample treatment in a 30°C waterbath, this step does not appear to stimulate an increase in activity in the treatment subsamples that would otherwise bias our results. Treatment samples, in fact, yielded lower calculated hydrolysis rates (due to improvements in substrate recovery) than no-treatment controls, which are processed immediately without incubation in the waterbath (Fig 1.3c). The general activity patterns in the chromatograms, which may be summarized by how quickly the fluorescence in the low molecular weight portion of the chromatogram increased over time, were similar between no treatment and treatment controls despite differences in chromatogram quality and intensity (Fig 1.2). The relative rate of increase in fluorescence over time of low molecular weight substrate products within a particular incubation remained the same in the no treatment and treatment conditions (result of paired t-test, $P=0.98$), even in highly active sediments.

3.2 Competitive desorption effects on substrate recovery and calculated hydrolysis rates

Desorption treatment increased total integrated fluorescence intensity of the resultant chromatogram by a median of 66% ($P<0.001$) relative to a no-treatment control (Fig 1.3a), with a median increase in fluorescence of 8.3×10^6 mV ($P<0.001$). The improvement in fluorescence intensity is observed in both the high- and low-molecular weight portion of the chromatogram, but is particularly effective in improving recovery of the high molecular weight portion (Fig 1.3b). Recovery of high molecular weight substrate products is improved by a median of 200% ($P=0.01$), while recovery of low molecular weight substrate products is improved by a median of 39% ($P<0.001$).

The improved recovery of the substrate from the subsample results in a higher relative proportion of high- to low-molecular weight substrate than is observed in the no-treatment controls. The desorption treatment therefore results in a lower calculated hydrolysis rate in treatment samples (Fig 1.3c), with a median decrease in maximum hydrolysis rate of 5 nM/hr in treatment subsamples relative to no-treatment controls ($P<0.01$).

The competitive desorption treatment improves substrate recovery in all sediments and substrates tested (Fig 1.4), although there is some variation in the percent improvement dependent upon sampling site (Fig 1.4a) and substrate (Fig 1.4b).

3.3 Comparison of the competitive desorption treatment with low pH-extraction

Several extraction treatment methods were tested in three preliminary experiments in order to develop and optimize the desorption protocol. Competitive desorption, using 700μM unlabeled substrate, was compared to alternative extraction treatments using solutions with pH of either 10 or 11, as well as a no treatment control (Fig 1.5a). Competitive desorption was found to be most effective at recovering chondroitin, increasing integrated fluorescence intensity by a median of 66% ($P<0.01$), while both pH extraction treatments actually decreased substrate recovery. Based on these results, competitive desorption was chosen as the basis of the extraction treatment method.

The concentration of unlabeled substrate to use during competitive desorption treatment was optimized in a second experiment (Fig 1.5b), comparing 700μM, 1400μM, and 2800μM unlabeled substrate concentrations to a no-treatment control. 2800μM concentrations improved integrated fluorescence intensity by a median of 32% ($P<0.01$), a large improvement over 700μM at 20% ($P=0.06$) and a slight improvement over 1400μM concentrations at 28%

(P=0.07). Based on these results, competitive desorption with 2800µM concentrations of unlabeled substrate was chosen as the basis of the final extraction treatment protocol.

The addition of SDS to competitive desorption was compared to a no-treatment control in a third experiment (Fig 1.5c) to determine whether the inclusion of SDS with competitive desorption provides additional substrate recovery. While competitive desorption both with and without SDS improved substrate recovery, the addition of SDS increased integrated fluorescence intensity by a median of 32% ($P<0.001$), whereas competitive desorption without SDS yielded a 17% increase ($P<0.001$). Therefore, competitive desorption with SDS was chosen as the final extraction treatment protocol.

3.4 Applicability of competitive desorption treatment using multiple substrates in sediments from diverse environments

We measured extracellular enzymatic hydrolysis in sediments from varied settings, and were able to detect hydrolysis at activities ranging from near-zero in the Eastern Mediterranean to more than 200nM/hr in parts of Guaymas Basin (Fig 1.6). The desorption treatment was more effective at increasing total integrated fluorescence intensities than a no-treatment control at every site (Fig 1.4a), with median percent improvement in total integrated fluorescence ranging from 7% (Guaymas core P1 depth 55cm, P=0.02) to 200% (Mediterranean non-sapropel, depth 365cm, $P<0.001$).

The greatest improvements in total integrated fluorescence intensities were observed in the sediments where the no treatment controls were uninterpretable due to very low substrate recovery. One such example is Guaymas core P13 at 55cm (Fig 1.2b), with a median improvement in integrated fluorescence intensities of 190%. The high quality chromatograms observed after desorption treatment in these cases demonstrates that this protocol can expand the range of sediments in which enzymatic activities can be measured.

Even in cases where the median improvement in fluorescence was relatively minor, for example Guaymas core P1 at depth 55cm, the treatment often improved the chromatogram quality in other tangible ways (Fig 1.2c). Guaymas P1-55cm exhibited very high hydrolytic activity, but the treatment recovered a large portion of high molecular weight substrate that was not recovered in the no treatment control, such that the final calculated hydrolysis rate was 110.6 nM/hr in the treatment incubation relative to the much higher 169.8 nM/hr in the no-treatment control.

The desorption treatment was effective in improving total integrated fluorescence intensity for both chondroitin and laminarin (Fig 1.4b, e.g. Fig 1.2). The treatment had a greater effect on laminarin recovery than chondroitin recovery: laminarin integrated fluorescence intensity was improved by a median of 140% ($P<0.001$), while chondroitin fluorescence was improved by a median of 20% ($P=0.01$).

4 Discussion

Microbial communities in sediments play an important role in driving key biogeochemical cycles. Organic carbon cycling is often the dominant metabolic function of microbial communities in subsurface environments (e.g. Biddle et al., 2006; Fry et al., 2008; Lloyd et al., 2013). Reliable measurements of enzymatic activities in sediments are fundamental to our understanding of microbial carbon-cycling potential in sedimentary environments. However, our ability to measure heterotrophic enzymatic activities directly in sediments, particularly subsurface sediments, has been hampered by the tendency for amended substrates to sorb to the sediment matrix. The extraction treatment presented here facilitates the measurement of enzyme activities in sediments by improving recovery of fluorescently labeled substrates in sediments from a range of geochemical settings. This treatment further enables the measurement of enzyme activities in sediments that might not otherwise yield usable data due to effects of sorption. This approach can be used to directly link microbial potential activities to genetic potential or biogeochemical processes, to better understand the role of microbial communities in subsurface carbon cycling. This approach may also be useful in remediation applications, where, for example, one would like to quantify the bioavailability of sediment-sorbed organic contaminants (e.g. Alexander, 2000; Megharaj et al., 2011), or in agricultural applications where the rate of recycling of nitrogen- or phosphorous-containing organic substrates by soil microbial communities is of interest for applications in optimizing food production (Berg, 2009) or minimizing fertilizer use (Adesemoye and Kloepper, 2009).

Treatment of sediment-sorbed fluorescently labeled substrates using competitive desorption and SDS proved effective in all sediments and substrates tested (Fig 1.4) improving chromatogram fluorescence intensities (Fig 1.3a) and leading to hydrolysis rate measurements reflecting improved recovery on high and low molecular weight substrate products (Fig 1.3b). Both highly active and near-zero activities were detected using this procedure (Fig 1.6) and desorption treatment improved chromatogram characteristics in both types of sediments (e.g. Fig 1.2a vs c). The geochemical and environmental contexts of the source sediments used to test the extraction treatment were varied, encompassing sapropelic sediments from the Mediterranean, with high concentrations of highly recalcitrant organic carbon; non-sapropelic, oligotrophic Mediterranean sediments; and sediments from Guaymas Basin ranging from highly compacted and sulfidic, to hemipelagic and diatom-rich, to coarse and sandy terrestrially-influenced sediments. The two substrates tested were distinct polysaccharides with distinct compositions and conformations, but both yielded improved recoveries when treated with the competitive desorption treatment (Fig 1.4b). The rates obtained by the competitive desorption treatment were not significantly affected by the incubation at 30°C in extraction buffer for 2 hours, while the improvements in chromatogram quality and hydrolysis rate calculations were substantial.

The hydrolysis rates measured in subsurface sediments highlighted the contrasting potential activities of sediment microbial communities in the Gulf of Mexico and Mediterranean Sea. Most locations in Guaymas Basin were much more active (ca. 100-200 nM/hr) than in the Eastern Mediterranean (ca. 0-20 nM/hr, Fig 1.6). Both of these subsurface sites exhibited lower activities than have been observed in surficial sediments (ca. upper 15cm of sediments) in previous studies, including Arctic sites (Arnosti, 2008) and sediments from the Gulf of Mexico (Arnosti et al., 2009). While the effect of desorption treatment on producing lower hydrolysis rates may have contributed to the difference in hydrolysis rates observed in this study, the 2-4 order of magnitude difference in microbial community abundance between shallow surface sediments and deeper subsurface sediments (Kallmeyer et al., 2012) likely underlies the considerable difference in measured rates. The activities measured here are potential enzymatic activities, and thus reflect the relative capacities of the nascent microbial communities to access the added substrate. The difference in rates observed between Guaymas and Mediterranean sediments suggests that these lower rates may also be indicative of differences in microbial communities between surficial and subsurface depths.

The efficacy of the extraction treatment in all of these settings suggests that the competitive desorption approach may be useful as a general (and therefore standardizable) approach to substrate recovery in sediment enzyme activity measurements. The competitive desorption treatment was especially effective at recovering the high molecular weight fraction of substrate products (Fig 1.3b). Recovery of high molecular weight substrate products is particularly useful in natural settings, because high molecular weight compounds are more likely to sorb to sediment than low molecular weight compounds (Podoll et al., 1987). The rates calculated from treatment samples may better reflect the potential hydrolysis rate occurring in the incubation, whereas the higher rate calculated for the untreated sample may be exaggerated due to disproportional sorption of the high molecular weight fraction. The demonstrated applicability of this treatment in sediments from broad environmental settings may be due to the mechanism of competitive desorption, which directly competes for sorption sites with a substrate of interest, regardless of the complex combination of factors that may affect adsorption rates (Pignatello and Xing, 1996) that may vary widely depending on the sediment and substrate characteristics.

Limitations of this extraction treatment will therefore most likely occur when sorption is more irreversible, i.e. where the decrease in entropy due to complexation of a substrate with sediment is prohibitively large. In this case, once sorbed a substrate is less likely to exchange with the aqueous environment, leaving less opportunity for an unlabeled 'competitive desorber' to replace it and release the labeled substrate. Although the extraction treatment was effective in all settings, the overall percent improvement in total integrated fluorescence intensities was variable across sites (Fig 1.4a). For example, in the Eastern Mediterranean core substrate recovery was much better in the non-sapropel ("N" segments) than the sapropel ("S") segments, perhaps due to the high concentrations of organic carbon in the sapropel segments. Future experiments may test a wider range of substrates, and additional sediments and soils with geological histories and contexts not investigated in this study, to better estimate the true variance of this extraction treatment approach and its overall applicability.

This method has been developed to improve measurements made with high molecular weight fluorescent substrates prepared after the method of Arnosti (1996, 2003), and is focused on improving recovery of hydrolyzed fragments of high molecular weight substrates. Other common methods of measuring enzymatic activities include use of low molecular weight substrate proxies, typically consisting of monomers linked to MUF or MCA fluorophores, after the method of Hoppe (1983). Measurement of enzyme activities with substrate proxies rely on the release of a fluorophore that becomes fluorescent upon hydrolysis from the attached monomer. The method is thus affected by sorption of the freed fluorophore to the sediment matrix. In this case, an adaptation of the competitive desorption treatment would require use of a non-fluorescent analog of the MCA or MUF fluorophore.

Measuring the rate at which microbial communities hydrolyze organic compounds in subsurface sediments is essential to our understanding of subsurface ecosystems and their influence on biogeochemistry, environmental remediation, and agricultural productivity. The method presented here provides a promising means to more reliably and accurately measure heterotrophic extracellular enzymatic activities in sediments not otherwise amenable to these measurements.

5 Author Contributions

AH and CA designed the experiments. AH and RS set up and subsampled sediment incubations, and processed subsamples. AH analysed data. AH, CA, and RS wrote the manuscript.

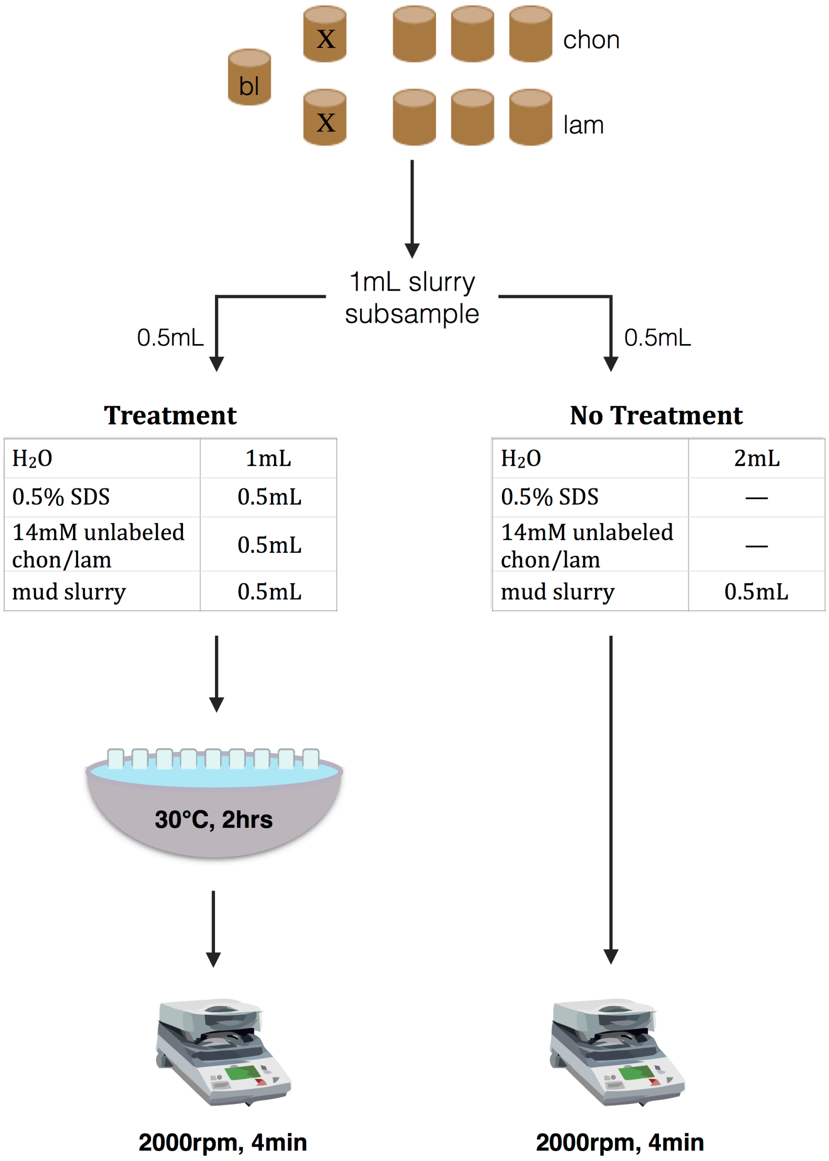
6 Funding

This work was supported by a CDEBI Graduate Fellowship to AH and a CDEBI research grant to CA, NSF award number OCE-0939564. Additional grants from NSF OCE-1449604 (to Andreas Teske) for sample collection in the Guaymas Basin, and NSF OCE-1332881 (to CA) provided partial support of this work. The collection (by Andreas Teske) of sediments from the Eastern Mediterranean and Marmara Sea was made possible by funding from the DARCLIFE project to Andreas Teske (UNC) and Kai Hinrichs (MARUM, Bremen).

7 Acknowledgments

The authors would like to thank Andreas Teske and Luke McKay who collected and generously provided the sediments used in this study. This publication is CDEBI Contribution 355.

8 Figures



Treatment

H_2O	1mL
0.5% SDS	0.5mL
14mM unlabeled chon/lam	0.5mL
mud slurry	0.5mL

No Treatment

H_2O	2mL
0.5% SDS	—
14mM unlabeled chon/lam	—
mud slurry	0.5mL

Fig 1.1 – Conceptual figure of extraction treatment protocol for competitive desorption with SDS. Nine incubations were conducted for each sediment section: one live blank ('bl'), and for each substrate one kill control ('X') and three live incubations. At subsampling, 1mL of slurry is removed from each incubation; 0.5mL of that subsample is treated with competitive desorption treatment (left), and 0.5mL receives no treatment (right). After centrifuging both treatments are syringe filtered and stored at -20°C.

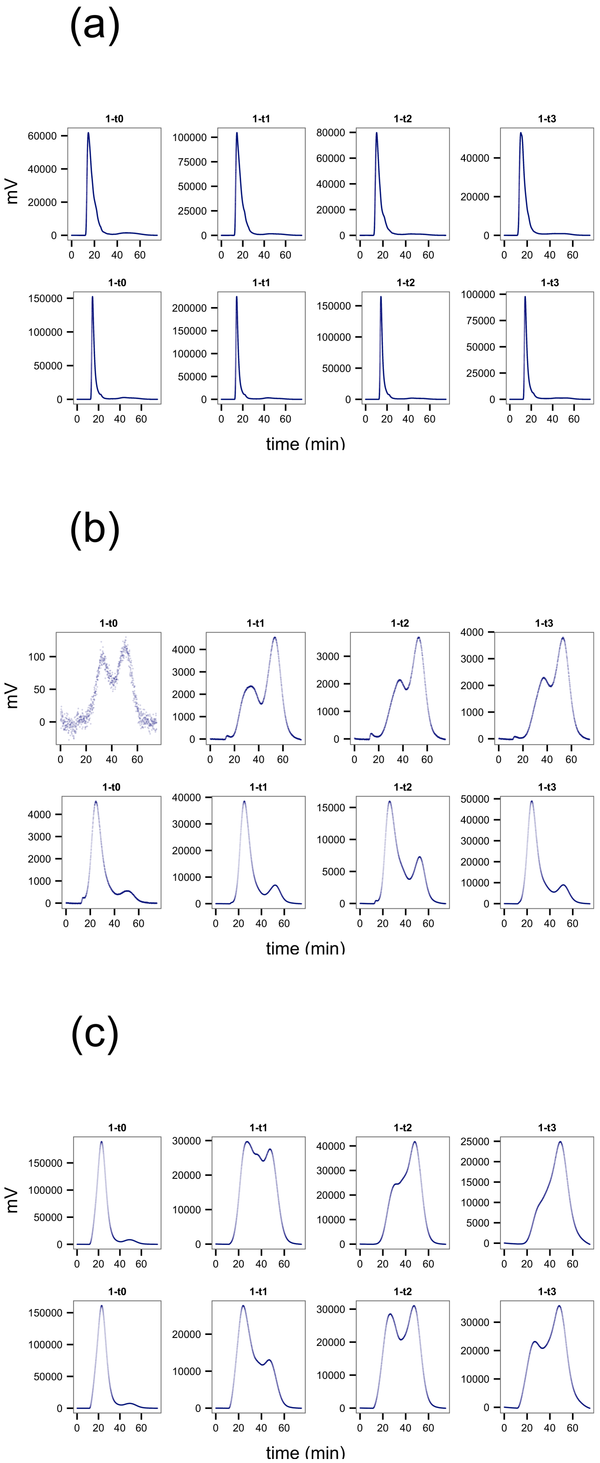


Fig 1.2 – Representative chromatograms comparing treatment (bottom row) to no treatment controls (top row) in one of the live replicate incubations for (a) Mediterranean sapropel S4, 385cm, chondroitin incubations (b) Guaymas core P13, 55cm, laminarin, and (c) Guaymas core P1, 55cm, laminarin. Note differences in scales on y axes. Improved chromatogram quality is seen in treatment incubations, with narrower peak widths, higher total integrated fluorescence, and a higher proportion of high- to low-molecular-weight substrate.

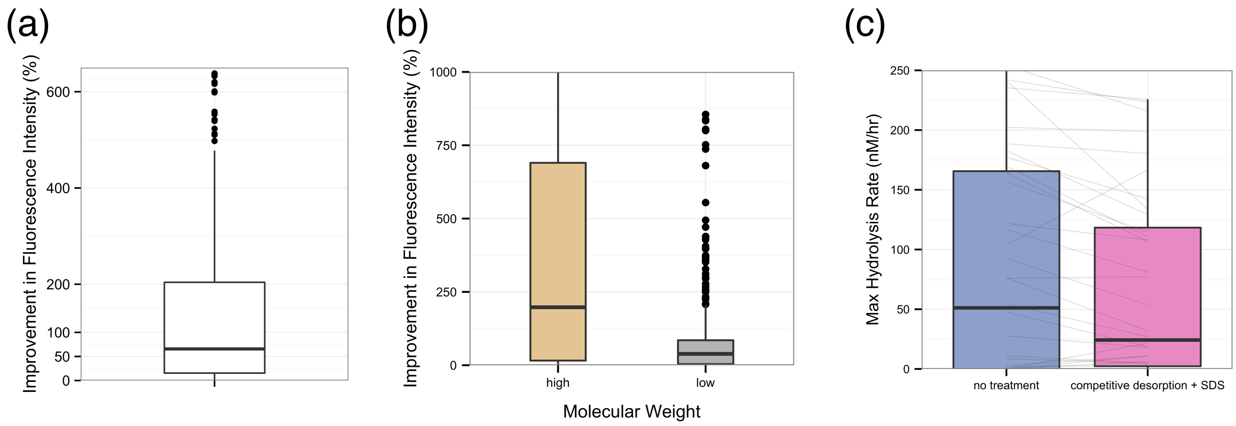


Fig 1.3 – Overall improvement in fluorescence intensity by competitive desorption and SDS treatment, for all experiments in Guaymas Basin and Eastern Mediterranean sediment. (a) The percent improvement in total integrated fluorescence intensity in desorption-treated samples relative to their no-treatment controls. (b) Percent improvement in total integrated fluorescence intensity using desorption treatment for the high- and low- molecular weight portions of each subsample. High molecular weight is operationally identified as the first third of a chromatogram, while low molecular weight is the last third. (c) Change in calculated maximum hydrolysis rate (nM/hr) in treatment versus no-treatment controls. Grey lines connect treatment and no-treatment subsamples from a single incubation.

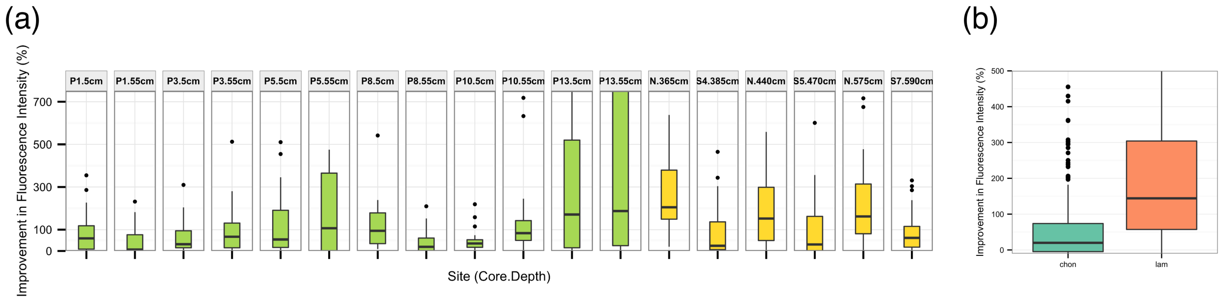


Fig 1.4 – Percent improvement in total integrated fluorescence intensity using desorption treatment method (relative to no-treatment control) (a) for each sediment section tested. Sediments from Guaymas Basin in green, Eastern Mediterranean in yellow, and (b) for each substrate tested.

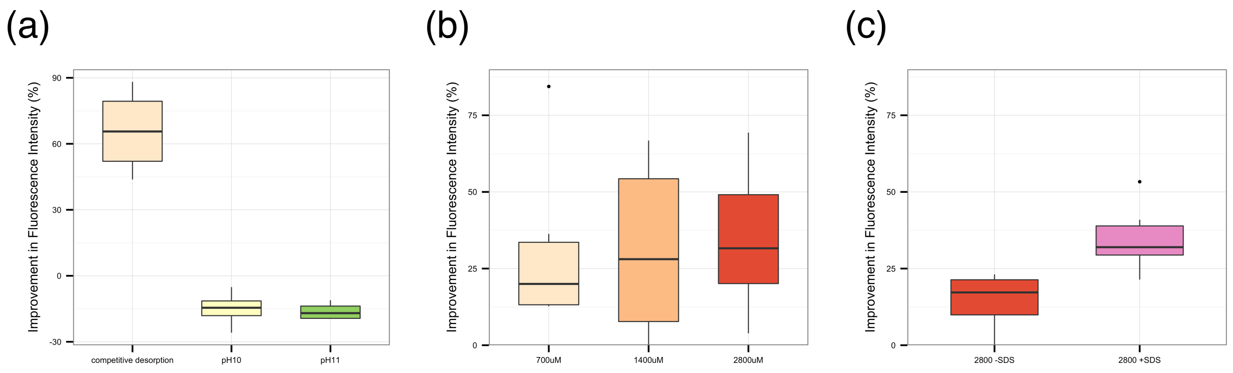


Fig 1.5 – Percent improvement in total integrated fluorescence intensity for alternative extraction treatment candidates relative to a no-treatment control for (a) PreX1, in which competitive desorption, pH=10, and pH=11 extraction conditions are tested, (b) PreX2, in which three candidate concentrations of unlabeled substrate (700μM, 1400μM, and 2800μM) in the extraction treatment are compared; and (c) PreX3, in which the presence or absence of SDS in competitive desorption are compared.

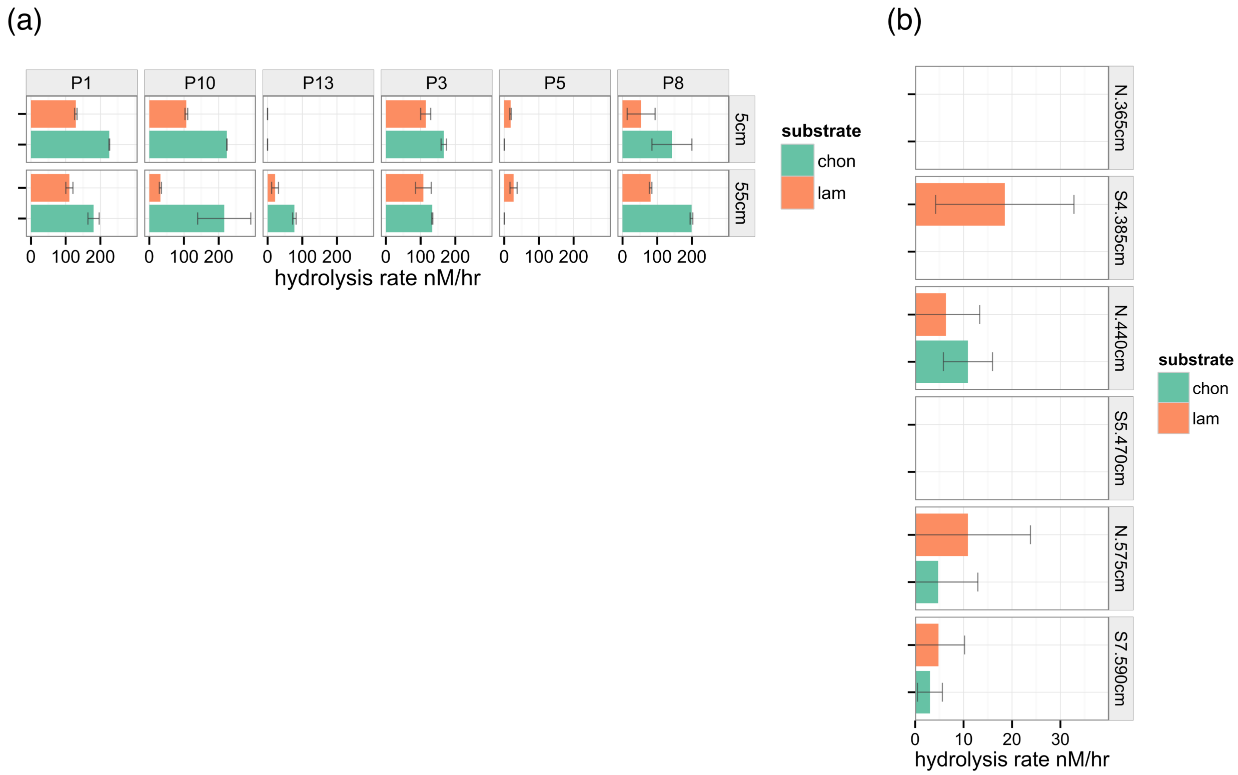


Fig 1.6 – Maximum hydrolysis rate (nM/hr) measured in sediments from (a) Guaymas basin and (b) Eastern Mediterranean. Note differences in scale on y axes.

REFERENCES

Adesemoye, A., and Kloepper, J. (2009). Plant microbe interactions in enhanced fertilizer use efficiency. *Appl. Microbiol. Biotechnol.* 85, 1–12.

Alderkamp, A.-C., van Rijssel, M., and Bolhuis, H. (2007). Characterization of marine bacteria and the activity of their enzyme systems involved in degradation of the algal storage glucan laminarin. *FEMS Microbiol. Ecol.* 59, 108–17. doi:10.1111/j.1574-6941.2006.00219.x.

Alexander, M. (2000). Aging, bioavailability, and overestimation of risk from environmental pollutants. *Environ. Sci. Technol.* 34, 4259–4265. doi:10.1021/es001069+.

Arnosti, C. (1996). A new method for measuring polysaccharide hydrolysis rates in marine environments. *Org. Geochem.* 25, 105–115. doi:10.1016/S0146-6380(96)00112-X.

Arnosti, C. (2003). Fluorescent derivatization of polysaccharides and carbohydrate-containing biopolymers for measurement of enzyme activities in complex media. *J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci.* 793, 181–91.

Arnosti, C. (2008). Functional differences between Arctic seawater and sedimentary microbial communities: contrasts in microbial hydrolysis of complex substrates. *FEMS Microbiol. Ecol.* 66, 343–51. doi:10.1111/j.1574-6941.2008.00587.x.

Arnosti, C., Ziervogel, K., Ocampo, L., and Ghobrial, S. (2009). Enzyme activities in the water column and in shallow permeable sediments from the northeastern Gulf of Mexico. *Estuar. Coast. Shelf Sci.* 84, 202–208. doi:10.1016/j.ecss.2009.06.018.

Berg, G. (2009). Plant-microbe interactions promoting plant growth and health: Perspectives for controlled use of microorganisms in agriculture. *Appl. Microbiol. Biotechnol.* 84, 11–18. doi:10.1007/s00253-009-2092-7.

Biddle, J. F., Lipp, J. S., Lever, M. a, Lloyd, K. G., Sørensen, K. B., Anderson, R., et al. (2006). Heterotrophic Archaea dominate sedimentary subsurface ecosystems off Peru. *Proc. Natl. Acad. Sci. U. S. A.* 103, 3846–51. doi:10.1073/pnas.0600035103.

Buckley, A., McKay, L. J., Turner, T., Chanton, J., Hensen, C., Benninger, L., et al. (2015). Biogeochemical and Microbial Survey of Gravity Cores from the Guaymas Basin and Sonora Margin. Talk and abstract. *Fall Meet. Am. Geophys. Union, San Fr. Dec. 14-18.*

Calvert, S., and Fontugne, M. (2001). On the late Pleistocene-Holocene sapropel record of climatic and oceanographic variability in the eastern Mediterranean. *Paleoceanography* 16, 78–94.

Chenu, C., and Stotzky, G. (2002). Interactions between microorganisms and soil particles: An overview. *Interact. between soil Part. Microorg. - Impact Terr. Ecosyst.*, 3–40.

Coolen, M. J. L., Cypionka, H., Sass, A. M., Sass, H., and Overmann, J. (2002). Ongoing modification of Mediterranean Pleistocene sapropels mediated by prokaryotes. *Science* 296, 2407–10. doi:10.1126/science.1071893.

Coolen, M. J. L., and Overmann, J. (2000). Functional Exoenzymes as Indicators of Metabolically Active Bacteria in 124,000-Year-Old Sapropel Layers of the Eastern Mediterranean Sea. *Appl. Environ. Microbiol.* 66, 2589–2598. doi:10.1128/AEM.66.6.2589-2598.2000.

Fry, J. C., Parkes, R. J., Cragg, B. A., Weightman, A. J., and Webster, G. (2008). Prokaryotic biodiversity and activity in the deep subseafloor biosphere. *FEMS Microbiol. Ecol.* 66, 181–196. doi:10.1111/j.1574-6941.2008.00566.x.

Gu, B. H., Schmitt, J., Chen, Z. H., Liang, L. Y., and McCarthy, J. F. (1994). Adsorption and Desorption of Natural Organic-Matter on Iron-Oxide - Mechanisms and Models. *Environ. Sci. Technol.* 28, 38–46. doi:10.1021/es00050a007.

Gu, B., Mehlhorn, T. L., Liang, L., and McCarthy, J. F. (1996). Competitive adsorption, displacement, and transport of organic matter on iron oxide: II. Displacement and transport. *Geochim. Cosmochim. Acta* 60, 2977–2992. doi:10.1016/0016-7037(96)00157-3.

Hmelo, L. R., Mincer, T., and Van Mooy, B. a S. (2011). Possible influence of bacterial quorum sensing on the hydrolysis of sinking particulate organic carbon in marine environments. *Environ. Microbiol. Rep.* 3, 682–8. doi:10.1111/j.1758-2229.2011.00281.x.

Hoarfrost, A. (2016). SedS. *Github Repos.* doi:10.5281/zenodo.233018.

Hoarfrost, A., and Arnosti, C. (2016). Investigating microbial activities driving organic matter transformations in the deep subsurface. *Biol. Chem. Oceanogr. Data Manag. Off.* Available at: http://www.bco-dmo.org/project/662055 [Accessed January 1, 2016].

Hoppe, H. (1983). Significance of exoenzymatic activities in the ecology of brackish water: measurements by means of methylumbelliferyl-substrates. *Mar. Ecol. Prog. Ser* 11, 299–308.

Kaiser, K., and Guggenberger, G. (2000). The role of DOM sorption to mineral surfaces in the preservation of organic matter in soils. *Org. Geochem.* 31, 711–725. doi:10.1016/S0146-6380(00)00046-2.

Kaiser, K., and Zech, W. (1999). Release of Natural Organic Matter Sorbed to Oxides and a Subsoil. *Soil Sci. Soc. Am. J.* 63, 1157. doi:10.2136/sssaj1999.6351157x.

Kallmeyer, J., Pockalny, R., Adhikari, R. R., Smith, D. C., and D'Hondt, S. (2012). Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc. Natl. Acad. Sci. U. S. A.* 109, 16213–6. doi:10.1073/pnas.1203849109.

Keil, R., Montluçon, D., Prahl, F., and Hedges, J. (1994). Sorptive preservation of labile organic matter in marine sediments. *Nature* 370, 549–552. doi:10.1038/370549a0.

Lloyd, K. G., Schreiber, L., Petersen, D. G., Kjeldsen, K. U., Lever, M. a, Steen, A. D., et al. (2013). Predominant archaea in marine sediments degrade detrital proteins. *Nature* 496, 215–8. doi:10.1038/nature12033.

Lutzow, M. V., Kogel-Knabner, I., Ekschmitt, K., Matzner, E., Guggenberger, G., Marschner, B., et al. (2006). Stabilization of organic matter in temperate soils: Mechanisms and their relevance under different soil conditions - A review. *Eur. J. Soil Sci.* 57, 426–445. doi:10.1111/j.1365-2389.2006.00809.x.

Megharaj, M., Ramakrishnan, B., Venkateswarlu, K., Sethunathan, N., and Naidu, R. (2011). Bioremediation approaches for organic pollutants: A critical perspective. *Environ. Int.* 37, 1362–1375. doi:10.1016/j.envint.2011.06.003.

Pignatello, J. J., and Xing, B. (1996). Mechanisms of slow sorption of organic chemicals to natural particles. *Environ. Sci. Technol.* 30, 1–11. doi:10.1021/es940683g.

Podoll, R. T., Irwin, K. C., and Brendlinger, S. (1987). Sorption of water-soluble oligomers on sediments. *Environ. Sci. Technol.* 21, 562–568. doi:10.1021/es00160a006.

Theng, B. K. G. (1979). Clay-Polymer Interactions: Summary and Perspectives. *Clays Clay Miner.* 30, 1–10. doi:10.1346/CCMN.1982.0300101.

Wegner, C.-E., Richter-Heitmann, T., Klindworth, A., Klockow, C., Richter, M., Achstetter, T., et al. (2013). Expression of sulfatases in Rhodopirellula baltica and the diversity of sulfatases in the genus Rhodopirellula. *Mar. Genomics* 9, 51–61. doi:10.1016/j.margen.2012.12.001.

Xing, P., Hahnke, R. L., Unfried, F., Markert, S., Huang, S., Barbeyron, T., et al. (2015). Niches of two polysaccharide-degrading Polaribacter isolates from the North Sea during a spring diatom bloom. *ISME J.* 9, 1410–1422. doi:10.1038/ismej.2014.225.

CHAPTER 2: RINGWATER INTRUSION ON THE MID-ATLANTIC BIGHT SHELF AFFECTS MICROBIALLY-DRIVEN CARBON CYCLING

A. Hoarfrost[1], JP Balmonte[1], S. Ghobrial[1], K. Ziervogel[3], J. Bane[1], G. Gawarkiewicz[2], C. Arnosti[1]

[1] Dept. of Marine Sciences, University of North Carolina – Chapel Hill

[2]Dept. of Physical Oceanography, Woods Hole Oceanographic Institution

[3]Institute for the Study of Earth, Oceans, and Space, University of New Hampshire

Corresponding author: Adrienne Hoarfrost (adrienne.l.hoarfrost@gmail.com)

In prep for *Limnology & Oceanography*

Running Head: Eddy Intrusion Effects on Carbon Cycling

Keywords

eddy intrusion, warm core ring, Mid Atlantic Bight, heterotrophy, carbon cycling, enzymatic activity

1. Introduction

Western boundary currents often influence an adjacent or nearby continental shelf. The Mid Atlantic Bight continental shelf in the northeastern U.S. in particular is affected by warm core rings originating from the Gulf Stream (e.g. Joyce et al. 1992; Gawarkiewicz et al. 2001; Chen et al. 2014; Zhang and Gawarkiewicz 2015). Warm core rings generally form on the north side of the Gulf Stream and drift to the north and west, where they eventually encounter the upper continental slope and the outer continental shelf. In recent years, there have been instances in which the path of the Gulf Stream jet has shifted well north of its normal meander envelope and is in close proximity to the shelf break and outer continental shelf (Gawarkiewicz et al. 2012; Ezer et al. 2013; Ullman et al. 2014).

Such shifts in ocean water circulation have the potential to profoundly affect the biological framework of life in the ocean. For example, upwelling of nutrient-rich deep water to the euphotic zone at the shelfbreak front in the Mid Atlantic Bight, where persistent upwelling contributes to enhanced primary productivity within the shelfbreak front and jet (He and Chen 2010; Zhang et al. 2013), fuels primary productivity in this region. Such interactions between physical and biological processes have long been understood as the foundation of the ocean's food web (Redfield 1958; McGillicuddy 2015). Understanding of finer-scale ocean circulation interactions with the shelf is less well-developed, in part because development of the instruments and capabilities to make higher resolution observations of relevant ocean parameters over sufficiently large temporal and spatial scales has occurred comparatively recently. In the last several years, however, new observational capabilities such as the Ocean Observatories Initiative Pioneer Array have highlighted increased exchange across the shelf break and the importance of warm core rings over the upper continental slope (Gawarkiewicz et al. 2018).

Recent studies have suggested that Gulf Stream influences over the continental shelf and slope south of New England have been increasing. Andres (2016) found that the initiation region for large amplitude Gulf Stream meanders has been shifting steadily westward since 1995 and large amplitude meanders are now occurring west of the New England Sea Mounts. Furthermore, the number of warm core rings formed annually has increased by roughly 50% for the time frame 2000-2016 compared to 1977-1999 (Monim 2017). Repeated cross-shelf glider transects from the Ocean Observatories Initiative Pioneer Array have shown that the mean salinity over the continental slope over a two year time period was 35.7 PSU, an increase of over 0.6 PSU relative to slope water mass properties from the 1970s and earlier (Gawarkiewicz et al., 2018).

Warm, salty waters intruding onto the shelf bring ecosystem changes: for example, new species have been documented on the continental shelf during seasons in which they are not normally present (Gawarkiewicz et al. 2018). However, the biogeochemical effects of such intrusions have not yet been investigated. The functional capabilities of microbial communities in ring features, and the timing and persistence of rings along the continental shelf, may affect the location and rate of carbon cycling along ocean basin margins. The activities of heterotrophic microbial communities are particularly important in this respect, since they are key drivers of carbon cycling, transforming and remineralizing organic matter, and generating new biomass. These processes function as a constraint on the amount of carbon recycled to the atmosphere as CO_2 or transported to deeper water depths (Azam and Malfatti 2007; Falkowski et al. 2008). The carbon-cycling capabilities of these communities is initiated by the activities of extracellular enzymes, which hydrolyze high molecular weight (HMW) organic matter into sizes sufficiently small to be transported into the cell. The assemblage of enzymes that a microbial community can produce affects the nature and quantity of organic substrates that microbial communities can access, as well as the rates at which they are hydrolyzed (Arnosti 2011). These functional patterns follow gradients across depth (Baltar et al. 2010b; Steen et al. 2012; Hoarfrost and Arnosti 2017), latitude (Arnosti et al. 2011), hydrographic properties (Baltar and Arístegui 2017; Hoarfrost and Arnosti 2017), and between coastal and open ocean regions (D'Ambrosio et al. 2014).

Distinct hydrolytic capacities of microbial communities within water masses result in different rates of carbon degradation in different regions of the ocean, often most obviously where boundaries between water masses are sharp (Baltar and Arístegui 2017). Gulf Stream warm core rings may transport a distinct microbial community with distinct hydrolytic capacities (Baltar et al. 2010a). These distinct functional capacities may affect carbon cycling over the Mid Atlantic Bight shelf break during intrusions of eddy-derived waters onto the continental shelf; however, this possibility has not yet been investigated.

Recent years have brought increased use of new sensors, gliders, and floats that provide continuous physical and chemical data, enabling high-resolution spatial and temporal tracking of specific water masses, including increased frequency of ring intrusions on the shelf (Andres 2016; Gawarkiewicz et al. 2018). In contrast, continuous and/or high-resolution measurements of microbial extracellular enzymatic activity – the initial step of carbon cycling – is not yet possible with similar spatial and temporal resolution. Although high-resolution automated collection of samples of microbial transcripts provide new insight into microbial dynamics (e.g. Otteson et al 2014; Aylward et al 2015), rates of processes cannot be inferred from such samples. Instead, here we focus on ship-based sampling and incubation of 'end points' of ring intrusion, by comparing microbial community activities in surface and bottom waters on a transect along the Mid Atlantic Bight shelf and shelf break during an eddy intrusion event. This transect included stations spanning shelf water, slope water, and ring-derived water. Measurement of microbial activities – potential activities of enzymes hydrolyzing peptides and polysaccharides, as well as bacterial protein production – across these distinct water masses yields insight into the biogeochemical capabilities of microbial communities associated with interactions between the Gulf Stream and the continental shelf on the Mid Atlantic Bight.

2. Methods

Hydrographic sampling was conducted with a SeaBird 911+ CTD from the R/V *Endeavor* between the 27th and 28th of April 2015 (cruise EN556). Vertical profiles were sampled at four stations between the 63 m and 207 m isobaths roughly along 71°W. Station 1 was located at 40.7071°N 71.028°W, Station 2 at 40.4622°N 71.0008°W, Station 3 at 40.3084°N 71.0048°W, and Station 4 at 40.0702°N 71.0052°W (Table 2.1, SI Fig 2.1). Water masses were identified (see Results) based on temperature and salinity characteristics, as well as the observation of warm core eddy dynamics from sea surface temperature satellite observations in this region during the time period of sampling.

2.1 Seawater collection

Seawater was collected from 1 meter below surface and within a few meters of the bottom at each station using a Niskin rosette equipped with a CTD sensor. Bottom sampling depths were 58 m (Stn. 1), 78 m (Stn. 2), 97 m (Stn. 3), and 199 m (Stn. 4; Table 2.1). Seawater was transferred to 20 L carboys that were rinsed three times with water from the sampling depth and then filled with seawater from a single Niskin bottle, using silicone tubing that had been acid washed then thoroughly rinsed with distilled water prior to use. From each carboy, water was dispensed into smaller glass containers that were cleaned and pre-rinsed three times with water from the carboy prior to dispensing. This water was used to measure bacterial productivity and the activities of polysaccharide hydrolases, peptidases, and glucosidases. A separate glass Duran bottle was filled with seawater from the carboy and sterilized in an autoclave for 20-30 minutes to serve as a killed control for microbial activity measurements.

2.2 Incubation setup and subsampling – polysaccharide hydrolases

The potential of the seawater microbial community to hydrolyze six high molecular weight polysaccharides (arabinogalactan, chondroitin sulfate, fucoidan, laminarin, pullulan, and xylan) was investigated in surface and bottom water at all four stations. These substrates were chosen for their diverse molecular structure, and because they are all found in the marine environment and/or enzymes able to target these substrates are present in marine microorganisms (e.g. Alderkamp et al., 2007; Martinez-Garcia et al., 2012; Wegner et al., 2013). Substrates were labeled with fluoresceinamine, after the method of Arnosti (1996, 2003).

For each substrate, three 50 mL falcon tubes were filled with seawater and one 50 mL falcon tube was filled with autoclaved seawater to serve as a killed control. Substrate was added at 3.5 μM monomer-equivalent concentrations, except for fucoidan, which was added at 5 μM concentrations (a higher concentration was necessary for sufficient fluorescence detection). Two 50 mL falcon tubes – one with seawater and one with autoclaved seawater – with no added substrate served as blank controls. Incubations were stored in the dark at as close to *in situ* temperature as possible, given the finite number of temperature-controlled incubators aboard ship (Table 2.1).

Subsamples of the incubations were collected at time zero, and at six subsequent timepoints (t1-t6): 2 days, 5 days, 10 days, 17 days, 30 days, and 42 days. These timepoints were chosen since it is impossible to know *a priori* at what timepoint hydrolytic activity can be detected. Here, we report the data from the first three timepoints (2, 5, and 10 days), since all activities that were detectable throughout the timecourse of incubation were detected by the 10 day timepoint. At each timepoint, 2 mL of seawater was collected from the 50 mL falcon tube using a sterile syringe, filtered through a 0.2 μm pore size syringe filter, and stored frozen until processing.

The hydrolysis of high molecular weight substrate to lower molecular weight hydrolysis products was measured using gel permeation chromatography with fluorescence detection, after the method of Arnosti (1996, 2003). In short, the subsample was injected onto a series of columns consisting of a 21 cm column of G50 and a 19 cm column of G75 Sephadex gel. The fluorescence of the column effluent was measured at excitation and emission wavelengths of 490 and 530 nm, respectively. Hydrolysis rates were calculated from the change in molecular weight distribution of the substrate over time, as described in detail in Arnosti (2003). The pairwise similarity of the spectra of polysaccharide substrates hydrolyzed among sampling sites was calculated using the Jaccard similarity metric, and the statistical significance of the differences in hydrolytic spectra among the shelf, slope, and eddy-intrusion water masses was evaluated with the PERMANOVA metric.

Scripts calculating hydrolysis rates and producing the figures depicted in this manuscript are available at the associated Github repository (Hoarfrost 2017).

2.3 Incubation setup and subsampling – peptidases and glucosidases

The hydrolysis of seven low molecular weight substrate proxies was measured in surface and bottom waters from all four stations. Two substrates, α-glucose and β-glucose linked to a 4-methylumbelliferyl (MUF) fluorophore, were used to measure glucosidase activities. Five substrates linked to a 7-amido-4-methyl coumarin (MCA) fluorophore, one amino acid – leucine – and four oligopeptides – the chymotrypsin substrates alanine-alanine-phenylalanine (AAF) and alanine-alanine-proline-phenylalanine (AAPF), and the trypsin substrates glutamine-alanine-arginine (QAR) and phenylalanine-serine-arginine (FSR) – were used to measure exo- and endo-acting peptidase activities, respectively. These substrates collectively are derived from two major classes of organic matter, carbohydrates (α-glucose and β-glucose) and amino acids that are constituents of proteins (leucine, QAR, FSR, AAF, and AAPF). Furthermore, these substrates are cleaved by enzymes that hydrolyze their substrates in two distinct patterns: exo-acting enzymes (cleaving end-terminus residues: α-glucose, β-glucose, and leucine) and endo-acting enzymes (mid-chain cleaving: QAR, FSR, AAF, and AAPF). As with the polysaccharide substrate incubations, these substrates were incubated at saturating concentrations and thus measure potential enzymatic activities.

Hydrolysis rates of the substrates were measured as an increase in fluorescence as the fluorophore was hydrolyzed from the substrate over time (as in Hoppe, 1983; Obayashi and Suzuki, 2005). Incubations with the seven low molecular weight substrates were set up in a 96-well plate. For each substrate, triplicate wells were filled with a total volume of 200 μL seawater for experimental incubations; triplicate wells were filled with 200 μL autoclaved seawater for killed control incubations. Substrate was added at saturating concentrations. A saturation curve was determined with surface water from each station to determine saturating concentrations of substrate. Saturation curve incubations were conducted with leucine and β-glucose substrates, and as saturating concentrations were found to be similar for both substrates, this concentration was used as the saturating concentration for all glucosidase and peptidase substrates. The saturating concentration was identified as the lowest tested concentration of substrate at which additional substrate did not yield higher rates of hydrolysis. Fluorescence was measured over 24 hours incubation time with a plate reader (TECAN spectrafluor plus; 360 nm excitation, 460 emission), with timepoints taken every 4-6 hours. Hydrolysis rates were calculated from the rate of increase of fluorescence in the incubation over time relative to a set of standards of known concentration of fluorophore. The similarity of the spectra of glucosidase and peptidase substrates hydrolyzed among sampling sites were calculated as with polysaccharide substrates, using the Jaccard similarity metric and testing statistical significance with PERMANOVA.

Scripts to calculate hydrolysis rates and produce the figures shown here are available in the associated Github repository (Hoarfrost 2017).

2.4 Bacterial productivity measurements

Bacterial protein production was measured via ^{3}H-leucine incorporation by heterotrophic bacteria using the cold trichloroacetic acid (TCA) and microcentrifuge extraction method (as in Kirchman, 2001). All work was performed aboard ship. In brief, triplicate live samples of 1.5 mL seawater as well as one 100% (w/v) TCA-killed control were incubated with 23 µL of L-[3,4,5-3H(N)]-Leucine (PerkinElmer, NET460250UC) for between 4 and 24 hours in the dark at as close to *in situ* temperature as possible. Live samples were then killed with 89 µL of 100% (w/v) TCA and centrifuged (10,000 rpm at 4°C for 10 min) to pelletize cell material. The supernatant liquid was removed and 1 mL of 5% (w/v) TCA solution was added, followed by vortex mixing and centrifugation. Supernatant removal, mixing, and centrifugation were repeated using 1 mL of 80% ethanol solution. Finally, the supernatant liquid was removed and each sample was dried overnight. After drying, 1 mL of scintillation cocktail (ScintiSafe 30% Cocktail, Fisher SX23-5) was added and incorporated radioactivity was measured using a LSA scintillation counter (PerkinElmer Tri-Carb 2910TR). Leucine incorporation rate was calculated from the incorporated radioactivity, compared to 1 mL of scintillation cocktail spiked with 23 μL of L-[3,4,5-3H(N)]-Leucine radioactivity, divided by incubation time.

2.5 Dissolved organic carbon measurements

Water samples for DOC measurements were collected from the Niskin bottles immediately after retrieval, before any other sampling took place. Clean and acid washed syringes, tubing, and filter holders were used for each sampling. Duplicate DOC samples were filtered using the same 60 cc syringe through combusted glass fiber filters (Whatman 1825-025) secured within a polycarbonate filter holder into two combusted 20 mL scintillation vials and acidified using 100 μL of 50% phosphoric acid, then immediately frozen at -20°C. DOC samples were analyzed by high temperature catalytic oxidation (HTCO) using a Shimadzu Total Organic Carbon analyzer (TOC-8000A/5050A).

3. Results

3.1 Water mass characteristics

The temperature/salinity properties from Stations 1-4 can be examined to determine the types of water masses present in April 2015 at the time of sampling. Overviews of T/S characteristics that define water masses in this region appear in Wright and Parker (1976) and Lentz (2003). Key water masses include shelf water beneath the thermocline, known as the Cold Pool, with temperatures typically 10°C or below, and salinities less than 34.0 PSU. Lentz (2003) has identified salinities between 34.0 and 35.0 PSU as shelfbreak frontal water. Slope water is generally 35.0-35.1 PSU. We define warm core ring water as having a salinity ≥ 35.5 PSU, although salinities within warm core rings can be 36.0 PSU or higher shortly after formation.

Temperature and salinity profiles from the CTD casts are shown in Fig 2.1, along with a T/S plot for the four stations and a satellite view of sea surface temperature on April 28, the day of sampling for Stns. 3 and 4.

The water mass properties from Stns. 1 and 2 were typical of shelf water Cold Pool masses in April, with temperatures below 7°C and salinities below 34 PSU. The upper water column from Stns. 3 and 4 also had temperature and salinities typical of shelf water farther offshore, although both temperature and salinity were slightly higher than at Stns. 1 and 2 (Fig 2.1). The salinity near the bottom in Stn. 3 was above 35 PSU, which is typical of slope water that includes mixtures of slope water and warm core ring water. The salinity for Stn. 4 was over 35.5 PSU from 120 m depth to the bottom of the profile at 200 m depth, indicating properties of a warm core ring water mass (Table 2.1, Fig 2.1). Dissolved organic carbon (DOC) was highest at Stns. 1 and 2 surface (120-135 µM), with decreasing concentrations offshore (80-90 µM at Stn. 2 bottom, Stn. 3, and Stn. 4 surface), and the lowest observed DOC concentrations at Stn. 4 bottom (63.5 µM) (Table 2.1).

Contemporary sea surface temperature observations during the time leading up to and during the dates sampled show the influence of a warm core eddy-derived warm filament intrusion on the shelf break where Stn 4 was sampled (Fig 2.1d, SI Fig 2.1). One warm core ring approached the survey line from the west during April 5 to 22, followed by a second warm core ring beginning on April 22. By April 27-28, the dates of sampling, the activity of these two rings appear to have "pushed" a warm water filament towards the survey line. The activity of these two rings likely brought warm/salty water to the upper continental slope at Stn. 4. The T-S character (~12.5°C, ~35.6 PSU) of the water below 120 m at Stn. 4 is consistent with recent observations of water within the Gulf Stream jet at the southern perimeter of the Mid Atlantic

Bight (SI Fig 2.2), suggesting the deep water at Stn. 4 is of Gulf Stream origin, likely transported to the survey area by warm core rings.

3.2 Microbial protein production

Bacterial productivity was particularly low in Stns. 3 and 4 bottom water, at 7 and 3 pmol L^{-1} h^{-1} respectively, but was considerably higher in surface water of the same stations, at 42 and 92 pmol L^{-1} h^{-1} respectively (Table 2.1). At Stns. 1 and 2, there was less difference between surface and bottom waters, and bacterial productivity was moderate compared to Stns. 3 and 4, ranging from 13 to 21 pmol L^{-1} h^{-1}.

3.3 Peptidase and glucosidase activities

Differences in the spectrum of peptide and glucose substrates hydrolyzed, as well as hydrolysis rates, were notable among shelf and shelf break water masses (Fig 2.2). The spectrum of peptide and glucose substrates hydrolyzed differed according to water mass source, resulting in significantly different Jaccard similarities of hydrolytic spectra between Stns. 1 and 2 Cold Pool shelf waters, Stn. 3 and 4 surface warm shelf waters, Stn. 3 bottom slope waters, and Stn. 4 bottom Gulf Stream eddy waters (PERMANOVA P=0.006). At Stns. 1 and 2, which corresponded with Cold Pool shelf waters, comparatively few substrates were hydrolyzed, and activities were quite low, ranging between 0 and 5.2 nmol L^{-1} h^{-1}. At these two stations, activities were dominated by hydrolysis of QAR-trypsin and either AAF-chymotrypsin (Stn. 2; Stn. 1 bottom water) or α-glucose (Stn. 1 surface water). Nearer the shelf break, Stns. 3 and 4 showed a much wider spectrum of activities and considerably higher hydrolysis rates. The surface warm shelf waters of Stns. 3 and 4, which had similar physical characteristics to eachother (Fig 2.1), also showed similar hydrolysis profiles, dominated by leucine amino peptidase activities averaging 79-88 nmol L^{-1} h^{-1}, with trypsin (QAR, FSR) and chymotrypsin (AAF, AAPF) activities ranging between 4 and 21 nmol L^{-1} h^{-1}, and very low but detectable levels of α- and β-glucosidase activity. Bottom waters of Stn. 3, corresponding to slope waters, showed lower leucine aminopeptidase activity (24.1 nmol L^{-1} h^{-1}) compared to surface water from the same station; activities of QAR-trypsin (at 24.1 nmol L^{-1} h^{-1}) were, however, higher than in surface water of Stn. 3. At Stn. 3 (both depths) as well as surface water of Stn. 4, all of the peptidase and glucosidase substrates were hydrolyzed. In Stn. 4 bottom water, corresponding to the Gulf Stream eddy water, however, a much narrower spectrum of activities was measured: no FSR, AAF, or α-glucose were hydrolyzed. This narrow spectrum of activities did not translate to low hydrolysis rates for the other substrates, however, since leucine aminopeptidase activity was 21.7 nmol L^{-1} h^{-1}, and AAPF-chymotrypsin and QAR-trypsin activities were 12.7 and 4.4 nmol L^{-1} h^{-1}, respectively (Fig 2.2).

3.4 Polysaccharide hydrolase activities

Polysaccharide hydrolase activities, measured initially after 48 hours' incubation, also showed distinct differences among stations and depths in the spectrum of substrates hydrolyzed, as well as in hydrolysis rates (Fig 2.3). These hydrolytic spectra were significantly different among Cold Pool shelf (Stns. 1 and 2), warm shelf break (Stns. 3 and 4 surface), slope (Stn. 3 bottom), and Gulf Stream derived eddy masses (Stn. 4 bottom, PERMANOVA P=0.048). Activities were considerably higher at Stn. 2 than at the other three stations (summed activities of 26.7 and 44.7 nmol L^{-1} h^{-1} for surface and bottom water, respectively), and were dominated by chondroitin hydrolysis. As with peptidase and glucosidase activities, Stns. 3 and 4 surface waters also had similar polysaccharide hydrolysis profiles: chondroitin, laminarin, pullulan, and xylan were hydrolyzed at both sites in similar proportions. Summed hydrolysis rates were lowest in bottom waters at Stns. 3 and 4 (1.7 and 5.0 nmol L^{-1} h^{-1} respectively), although their hydrolysis profiles differed: arabinogalactan, chondroitin, and laminarin were hydrolyzed in Stn. 3 bottom water, while fucoidan, laminarin, pullulan, and xylan were hydrolyzed in Stn. 4 bottom water. Fucoidan and arabinogalactan were only hydrolyzed in bottom waters, while pullulan was only hydrolyzed in surface waters and the bottom waters at Stn. 4.

Given the duration of the polysaccharide incubations, the timecourse of substrate hydrolysis provides additional information about the response of a microbial community to specific polysaccharides, since multi-day incubations allow sufficient time for growth responses, as well as enzyme induction. Rapid hydrolysis suggests that a large fraction of a microbial community can hydrolyze a substrate or that the active portion of the community is able to respond quickly and at high capacities, while hydrolysis that develops late in a timecourse indicates an activity carried out by a small or slow-growing fraction of the community. Substrate hydrolysis patterns and response times to the six polysaccharides evolved considerably over a 10-day incubation period across substrate, depth, and location (Fig 2.4). Over this time frame, hydrolysis of chondroitin and laminarin became evident in all incubations and generally became an increasing proportion of total polysaccharide hydrolysis over time. Pullulan was rapidly hydrolyzed in all surface waters, but only in bottom waters at Stn. 4. Arabinogalactan was only hydrolyzed in bottom waters, and only at Stns. 1 and 3.

4. Discussion

The frequency of warm core rings in the slope region south of New England has increased in recent years. In November 2009 (Ullman et al. 2014) and December 2011

(Gawarkiewicz et al., 2012), the North Wall of the Gulf Stream was in close proximity to the shelf break south of New England. Observations from the National Marine Fisheries Service, as part of the ECOsystems MONitoring program, have shown that ring water extended shoreward across most of the continental shelf in September 2014 (Gawarkiewicz et al., 2018). More recently, the Ocean Observatories Initiative Pioneer Array identified a large cross-shelf intrusion of ring water in January 2017. Our study identifies an eddy intrusion event on the continental slope during the time of sampling in April 2015. We detect distinct biogeochemical characteristics among water masses along the northern part of the Mid Atlantic Bight during this eddy intrusion event, and explore the effect of these events on microbially-driven carbon cycling in this region.

A pattern in microbial carbon cycling capacity unique to warm core eddy intrusions may have implications for carbon biogeochemistry within warm core eddies, and in particular on the biogeochemical cycling capacities of the continental shelf during eddy intrusion events. Across distinct water masses sampled from the continental shelf, shelf break, slope, and the Gulf Stream derived eddy (Fig 2.1), microbial communities demonstrated differences in the spectrum and rates of hydrolysis of a suite of high- and low-molecular-weight organic substrates, as well as in bacterial productivity, indicative of differences in carbon cycling capacities. Gulf Stream derived water in bottom water of Stn. 4 in particular was distinct from other water masses sampled at other stations and depths, and was characterized by particularly low bacterial productivity (Table 2.1), the capacity to hydrolyze fewer peptide substrates relative to surface waters at the same station (Fig 2.2), and distinct polysaccharide hydrolase activities (Fig 2.3) as well as peptide and glucose substrate spectra (Fig 2.2). The hydrolytic pattern in Stn. 4 bottom waters contrasted with the shelf break surface waters of Stns. 3 and 4, the slope waters of Stn. 3 bottom water, and the cold, shallow shelf waters of Stns. 1 and 2, each of which were characterized by hydrolytic spectra and hydrolysis rates significantly different from the other water masses. These water masses were also characterized by widely different rates of bacterial productivity, with particularly high bacterial productivity in shelf break surface waters, moderate productivity in cold shelf waters, and relatively low productivity in slope waters that nevertheless was more than twice the rate in warm core ring waters (Fig 2.1a; Table 2.1).

Differences in functional capacities were evident over long as well as short time-scales, as determined by incubations that reflect the capacity of the microbial community to respond to addition of specific polysaccharides. Even after 10 days of incubation, during which a significant growth response is possible, some polysaccharides were not hydrolyzed (Fig. 2.4), indicating that the microbial community as a whole lacked the ability to hydrolyze particular substrates, whether because they lacked the genes that encode the necessary enzymes or because such genes were not activated during the incubation time. Hydrolytic capacities in water from the warm core ring intrusion, and the timecourse over which different substrates were hydrolyzed most rapidly, were distinct from other sampling sites of shelf and shelf break waters (Fig 2.4). This observation suggests that the timing and persistence of eddy intrusions may also influence the spectrum of organic matter hydrolyzed by microbial communities *in situ*.

A previous investigation of microbial activities in the Mid Atlantic Bight that sampled waters originating from a Gulf Stream eddy in close proximity to the continental shelf (Bullock et al. 2015) indicates that the hydrolytic and bacterial productivity patterns observed in this study are a consistent feature of Gulf Stream derived eddies and intrusions. Bullock and colleagues (2015) found that the same spectrum of polysaccharides – pullulan, laminarin, xylan, and fucoidan – were hydrolyzed after 2 days of incubation as in the present study, and bacterial productivity was also notably low in waters derived from the Gulf Stream compared to the other water masses they sampled. Although not all of the peptidase substrates used in the current study were also measured in Bullock et al. (2015), activities of leucine aminopeptidase, α-glucosidase, and β-glucosidase showed similar patterns to Stn. 4 bottom water, with moderate leucine aminopeptidase rates (ca. 12 nmol L-1 h-1 compared to 21.7 nmol L-1 h-1 in this study) and very low α- and β-glucosidase rates. The similarity of hydrolytic rates and spectra, as well as the characteristically low bacterial productivity, in warm core ring waters across the current study as well as Bullock et al. (2015) suggests that water derived from warm core rings may have a consistent, distinct biogeochemical imprint on the continental shelf. Other studies comparing bacterial productivity in water from farther offshore or from warm core ring sampling sites to coastal or shallower water sites (Baltar et al. 2009, Alonso-Sáez et al. 2012) also found that bacterial productivity was lower in water originating from North Atlantic Central Water, consistent with both Bullock et al. (2015) and this study.

The warm core eddies driving eddy intrusion events on the Mid Atlantic Bight originate from the Gulf Stream and North Atlantic Central Waters of the Sargasso Sea. Eddies in the Sargasso Sea have strong effects on net community production, carbon export, and microbial community composition and biomass (Benitez-Nelson and McGillicuddy 2008; Ewart et al. 2008; Mouriño-Carballido 2009; Nelson et al. 2014) that mirror the patterns observed in eddy-driven intrusion events at our sampling sites. These studies demonstrate the distinct bacterial distributions, biomass, and productivity characteristic of eddies relative to surrounding water (Ewart et al. 2008; Mouriño-Carballido 2009), which are also associated with distinct microbial communities and bacterial productivity in the same range as that observed in this study (Nelson et al. 2014). Bacterial production The distinct community dynamics associated with eddy influences in the Sargasso Sea provide further evidence of the importance of physical mesoscale processes on the structure of bacterial distributions and activities; however, none of the aforementioned studies have measured enzyme activities, so the differences in hydrolytic spectra and rates in this study illuminate the potential net effect on the quality and quantity of carbon cycling during eddy intrusion events.

The hydrolytic patterns differentiating shelf, shelf break, and eddy intrusion water masses, and their biogeochemical consequences, are driven by more complex processes than a simple temperature relationship. Despite the significantly higher temperatures of eddy intrusion water at Stn. 4 bottom, this temperature difference does not correspond to higher rates of enzyme activity; moreover, bacterial productivity is lower than the surrounding shelf and shelf break water masses. The observed differences in hydrolysis rates and patterns are not due to temperature differences across sites, for either polysaccharide substrates (R^2=0.21. P=0.26) or peptide and glucose substrates (R^2=0.05, P=0.59); incubation temperatures in surface waters at all four stations were identical despite distinct differences in hydrolysis rates and capacities (Table 2.1, Fig 2.1), and the highest polysaccharide hydrolysis rates at 48 hours were at Stn. 2, which along with Stn. 1 were the coldest stations sampled (Fig 2.3). Moreover, hydrolysis rates of low molecular weight substrates in Stn. 4 bottom water were lower than in surface waters of Stns. 3 and 4, despite the fact that Stn. 4 bottom water had a higher *in situ* and incubation temperature (Table 2.1, Fig 2.1, Fig 2.2). Instead, we see a shift in the spectrum of substrates that are hydrolyzed, and hydrolysis rates that are independent of temperature, and may be more closely related to differences in functional capacities of the nascent microbial community. This suggests that biogeochemical models that rely on temperature as the primary factor driving biogeochemical carbon cycling, which would predict higher overall activities and bacterial production with increased warm core eddy intrusions on the continental shelf, would incorrectly predict biological impacts on carbon cycling rates.

There are several possible mechanisms for the observed differences in functional capabilities across water masses. Differences in functional capabilities across water masses may be partially driven by biogeographical differences in microbial communities, with eddy intrusions onto the continental shelf bringing with them a distinct microbial community with distinct hydrolytic capacities (Nelson et al. 2014). The differences in functional capacities observed across sites and water masses are in keeping with functional biogeographical patterns in substrate hydrolysis previously identified across latitude, station and depth (Arnosti et al. 2011; Hoarfrost and Arnosti 2017), and between on-shore and off-shore sites in the North Atlantic along similar scales of distance (D'Ambrosio et al. 2014). These patterns in functional biogeography mirror biogeographical patterns in microbial community composition (Fuhrman et al. 2008; Zinger et al. 2011; Ladau et al. 2013; Nelson et al. 2014). Evidence in support of a linkage between community composition and function includes an investigation from the East China Sea, where communities in different water masses from shallow and bottom water depths on the continental shelf exhibited differential abundances of genes involved in hydrolysis of starch and chitin-derived carbon sources (Wang et al. 2017). In the North Atlantic, moreover, the diversity of genes from glycosyl hydrolase family 5, one of the largest families of glycosyl hydrolases, differed significantly between the Mid Atlantic Bight and open ocean sites (Elifantz et al. 2008).

Differences in patterns of enzyme activities across water masses may also be due to differences in organic matter composition and/or primary producer communities. For example, the Mid Atlantic Bight shelf is typically dominated by diatoms and dinoflagellates (Falkowski et al. 1994), whereas the North Atlantic open ocean typically harbors a higher number of cyanobacteria than in coastal regions (Lomas and Bates 2004). These different taxa differ in organic matter compositions (Biersmith and Benner 1998), and thus may result in distinct water mass organic matter compositions that could also influence the activities of the nascent heterotrophic communities. Amendment with marine high molecular weight DOM is known to induce shifts in microbial community composition and expression of genes involved in carbon cycling (McCarren et al. 2010), while amendment with diatom- vs. cyanobacterial-derived dissolved organic matter (DOM) induces different microbial community responses in diversity and richness (Landa et al. 2014).

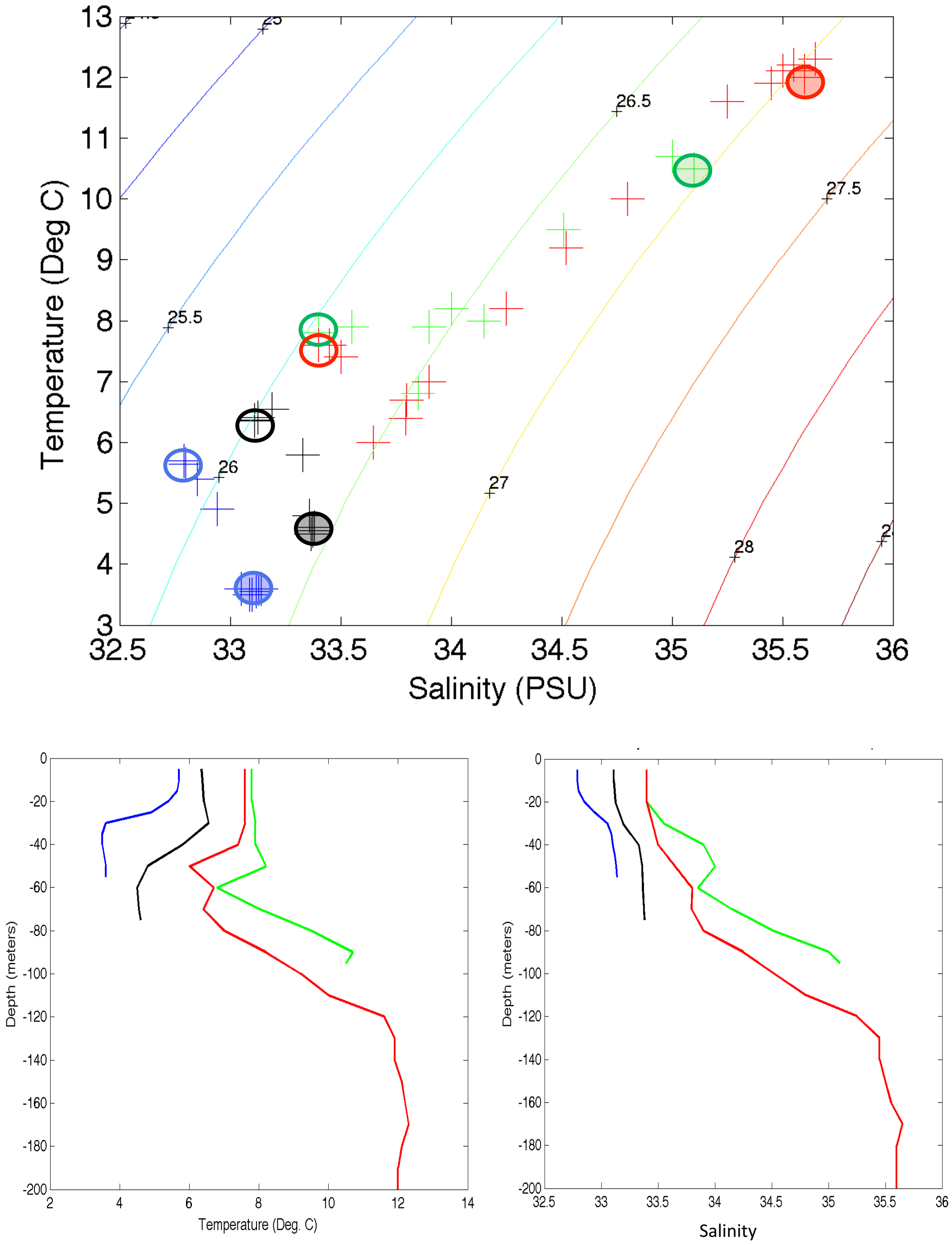
The distinct hydrolytic patterns in waters corresponding to the Gulf Stream eddy intrusion in Stn. 4 bottom water, as it contrasts with hydrolytic patterns in continental shelf and shelf break waters, highlights the potential biogeochemical importance and consequences of changes in Gulf Stream interactions with the continental shelf. Distinct hydrolytic activities and functional capacities within ring waters may alter the amount and quality of carbon cycling on the continental shelf. As Gulf Stream ring and meander dynamics shift in the coming years, these biogeochemical consequences are also likely to shift. The distinct hydrolytic spectra coupled with low bacterial productivity characteristic of eddy intrusion waters suggests both that the composition of organic matter remineralized on the Mid Atlantic Bight shelf break during eddy intrusion events is likely to shift, while the overall rate of remineralization is likely to be lower. As the frequency and duration of such eddy intrusion events increase in the future, this has implications for microbial carbon cycling on ocean margins on a large scale.

The frequency of ring intrusions onto the continental shelf, and the amount of time that such intrusions persist (e.g. Ullman et al., 2014; Zhang and Gawarkiewicz, 2015), may influence the hydrolytic capacities and microbial activities in these regions, affecting the quantity and composition of organic matter remineralized on ocean basin margins, and having a significant effect on biogeochemical cycling of carbon and nutrients. In the future, the frequency of ring intrusions are likely to increase, due to the recent destabilization of the Gulf Stream (Andres 2016) and the increasing number of warm core rings formed by the Gulf Stream (Monim 2017). As the continental shelf and slope south of New England is increasingly influenced by Gulf Stream waters via ring interactions and close contact with the north wall of the Gulf Stream (Gawarkiewicz et al. 2012), changes in microbial community function and biogeochemical carbon cycling driven by distinct functional capacities and patterns of bacterial productivity within warm core ring waters are also likely to shift.

Acknowledgements and Data

We thank the captain and crew of the R/V *Endeavor* for their able assistance at sea, and the members of the scientific party of EN556 for their help with the shipboard work. This project was funded by NSF (OCE-1332881 and -1736772 to CA; OCE-1657853 to GG). The authors declare no conflicts of interest. Raw data underlying hydrolysis rates, cell counts, bacterial productivity, nutrients, and CTD casts are hosted on the BCO-DMO repository under the project title “Latitudinal and depth-related contrasts in enzymatic capabilities of pelagic microbial communities: Predictable patterns in the ocean?”: http://www.bco-dmo.org/project/712359.

Figures



Temperature (Deg C)
Salinity (PSU)
13
12
11
10
9
8
7
6
5
4
3
32.5
33
33.5
34
34.5
35
35.5
36
25
25.5
26
26.5
27
27.5
28
Depth (meters)
Temperature (Deg. C)
0
-20
-40
-60
-80
-100
-120
-140
-160
-180
-200
2
4
6
8
10
12
14
Depth (meters)
Salinity
32.5
33
33.5
34
34.5
35
35.5
36

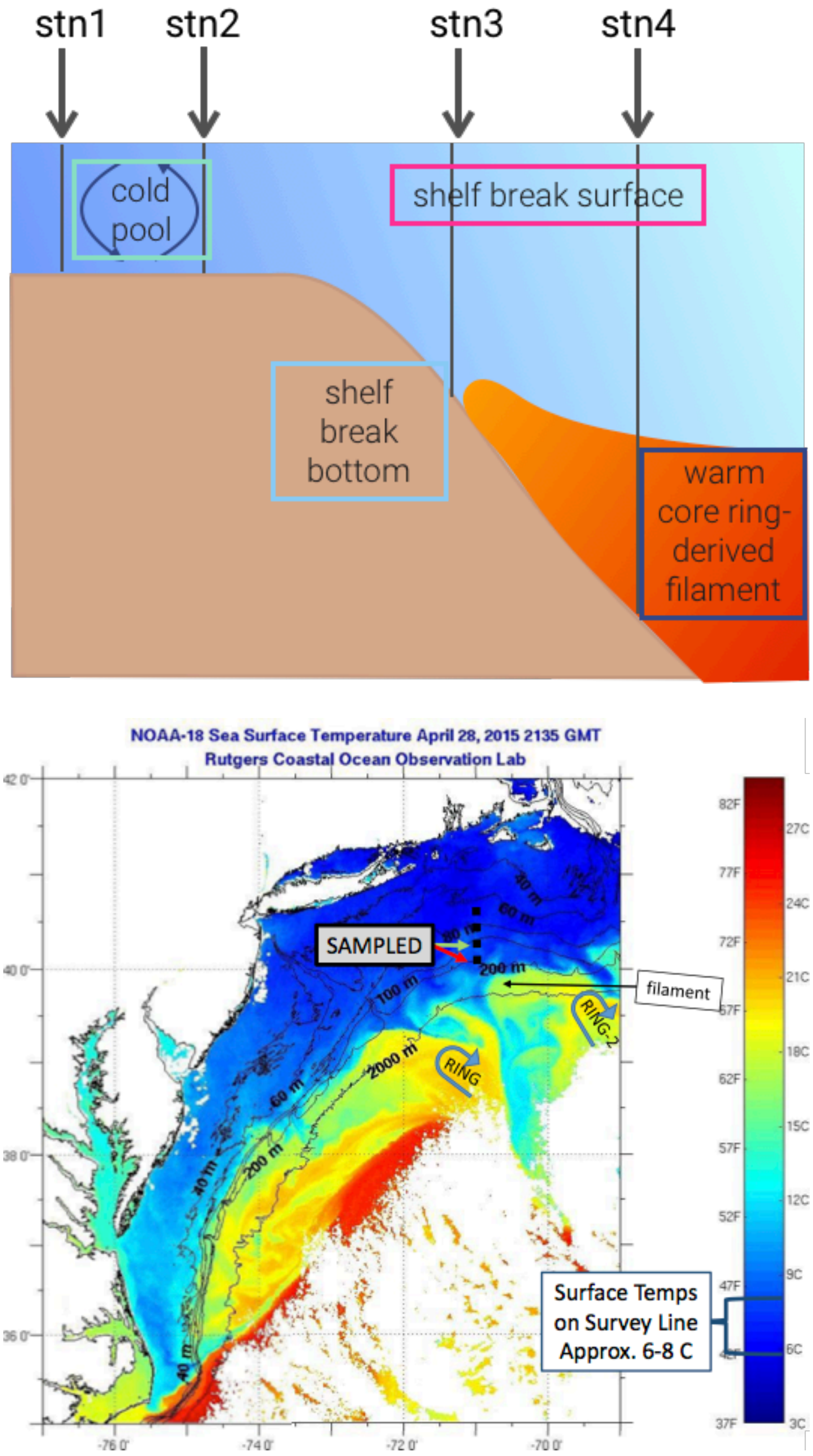


Figure 2.1 – **(a)** A temperature/salinity plot identifying water mass characteristics from the four stations. Sampling locations indicated by circles, filled circles bottom and empty circles surface

waters. **(b)** Temperature (left) and salinity (right) profiles at each station. Blue: Stn. 1; Black: Stn. 2; Green: Stn. 3; Red: Stn. 4. **(c)** Conceptual figure of water masses sampled. **(d)** Satellite imagery of sea surface temperature on April 28, the day Stns 3 and 4 were sampled. Stations surveyed are indicated by black squares. Two warm core rings near the survey sites are indicated by blue arrows, and a warm filament being pushed onto the continental shelf where Stn 4 bottom waters are sampled is indicated.

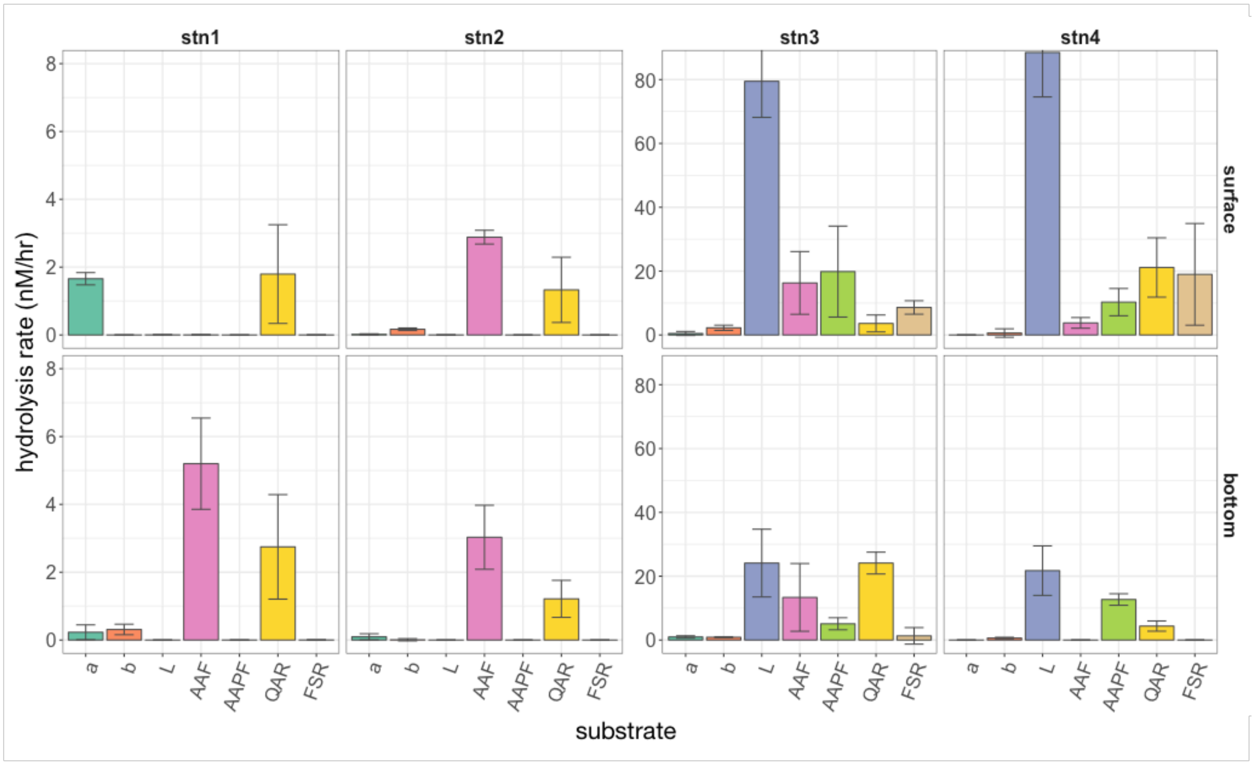


Figure 2.2 – Hydrolysis rate of the seven peptide and glucose substrates at each station (left to right) and surface and bottom (top and bottom panels). A-glu: α -glucose; B-glu: β -glucose; Leu: leucine; AAF: alanine-alanine-phenylalanine; AAPF: alanine-alanine-proline-phenylalanine; QAR: glutamine-alanine-arginine; FSR: phenylalanine-serine-arginine. Note the order of magnitude difference in y axes between Stns. 1 and 2 and Stns. 3 and 4.

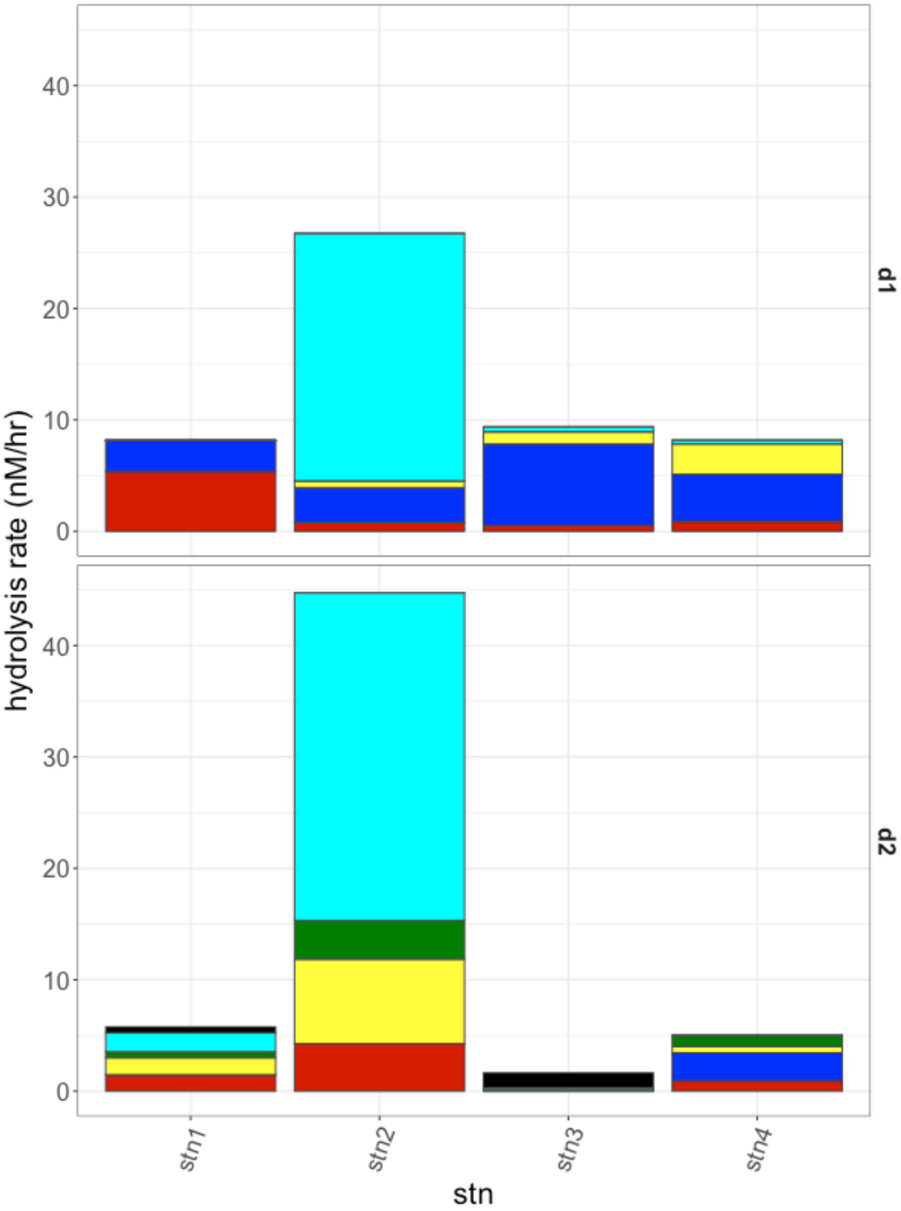


Figure 2.3 – Hydrolysis rates at the 48-hour sampling timepoint for the six polysaccharide substrates at each station (left to right) and surface and bottom (top and bottom panels). Black: arabinogalactan; Turquoise: chondroitin sulfate; Green: fucoidan; Yellow: laminarin; Blue: pullulan; Red: xylan.

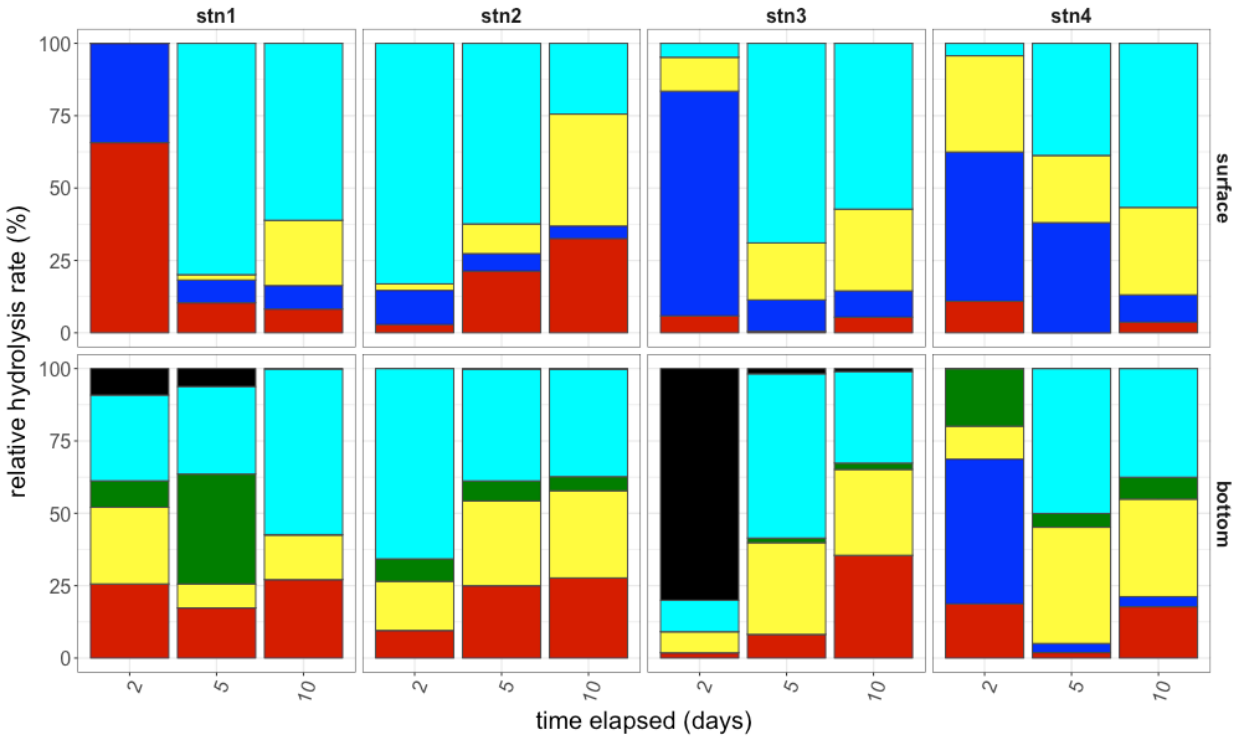


Figure 2.4 – Relative contributions of polysaccharide hydrolase activities to summed hydrolysis rates (y axis) at the time of subsampling (x axis) for each station (left to right) and depth (top and bottom). Black: arabinogalactan; Turquoise: chondroitin sulfate; Green: fucoidan; Yellow: laminarin; Blue: pullulan; Red: xylan.

Tables

	latitude (°N)	longitude (°E)	bottom depth (m)	samplin g depth (m)	sampling date (DD/MM/YYYY)	inc temp (°C)	*in situ* temp (°C)	*in situ* salinity (PSU)	*in situ* oxygen (ml L^{-1})	bact. prod. (pM/hr)	DOC (µM)
stn 1 surface	40.7071	-71.028	63	1	27/04/2015	8	5.7	32.8	7.2	13	135.0
stn 1 bottom	40.7071	-71.028	63	58	27/04/2015	4	3.6	33.1	6.7	20	137.1
stn 2 surface	40.4622	-71.0008	83.6	1	27/04/2015	8	6.3	33.1	7.1	21	120.1
stn2 bottom	40.4622	-71.0008	83.6	78	27/04/2015	4	4.7	33.4	6.5	21	81.0
stn3 surface	40.3084	-71.0048	102	1	28/04/2015	8	7.8	33.4	7.2	42	90.2
stn3 bottom	40.3084	-71.0048	102	97	28/04/2015	8	10.5	35.1	4.7	7	78.3
stn4 surface	40.0702	-71.0052	206.9	1	28/04/2015	8	7.7	33.4	7.3	92	89.0
stn4 bottom	40.0702	-71.0052	206.9	199	28/04/2015	14	12.0	35.6	4.7	3	63.5

Table 2.1 – Location, environmental data, bacterial production and dissolved organic carbon for each station and depth.

REFERENCES

Alderkamp, A.-C., M. van Rijssel, and H. Bolhuis. 2007. Characterization of marine bacteria and the activity of their enzyme systems involved in degradation of the algal storage glucan laminarin. FEMS Microbiol. Ecol. **59**: 108–17. doi:10.1111/j.1574-6941.2006.00219.x

Andres, M. 2016. On the recent destabilization of the Gulf Stream path downstream of Cape Hatteras. Geophys. Res. Lett. **43**. doi:10.1002/2016GL069966.Received

Arnosti, C. 1996. A new method for measuring polysaccharide hydrolysis rates in marine environments. Org. Geochem. **25**: 105–115. doi:10.1016/S0146-6380(96)00112-X

Arnosti, C. 2003. Fluorescent derivatization of polysaccharides and carbohydrate-containing biopolymers for measurement of enzyme activities in complex media. J. Chromatogr. B. Analyt. Technol. Biomed. Life Sci. **793**: 181–91.

Arnosti, C. 2011. Microbial extracellular enzymes and the marine carbon cycle. Ann. Rev. Mar. Sci. **3**: 401–425. doi:10.1146/annurev-marine-120709-142731

Arnosti, C., A. D. Steen, K. Ziervogel, S. Ghobrial, and W. H. Jeffrey. 2011. Latitudinal gradients in degradation of marine dissolved organic carbon. PLoS One **6**: e28900. doi:10.1371/journal.pone.0028900

Azam, F., and F. Malfatti. 2007. Microbial structuring of marine ecosystems. Nat. Rev. Microbiol. **5**: 782–91. doi:10.1038/nrmicro1747

Baltar, F., and J. Arístegui. 2017. Fronts at the Surface Ocean Can Shape Distinct Regions of Microbial Activity and Community Assemblages Down to the Bathypelagic Zone: The Azores Front as a Case Study. Front. Mar. Sci. **4**: 1–13. doi:10.3389/fmars.2017.00252

Baltar, F., J. Arístegui, J. M. Gasol, I. Lekunberri, and G. J. Herndl. 2010a. Mesoscale eddies: hotspots of prokaryotic activity and differential community structure in the ocean. ISME J. **4**: 975–988. doi:10.1038/ismej.2010.33

Baltar, F., J. Arístegui, J. Gasol, E. Sintes, H. van Aken, and G. Herndl. 2010b. High dissolved extracellular enzymatic activity in the deep central Atlantic Ocean. Aquat. Microb. Ecol. **58**: 287–302. doi:10.3354/ame01377

Benitez-Nelson, C. R., and D. J. McGillicuddy. 2008. Mesoscale physical-biological-biogeochemical linkages in the open ocean: An introduction to the results of the E-Flux and EDDIES programs. Deep. Res. Part II Top. Stud. Oceanogr. **55**: 1133–1138. doi:10.1016/j.dsr2.2008.03.001

Biersmith, A., and R. Benner. 1998. Carbohydrates in phytoplankton and freshly produced dissolved organic matter. Mar. Chem. **63**: 131–144. doi:10.1016/S0304-4203(98)00057-7

D'Ambrosio, L., K. Ziervogel, B. Macgregor, A. Teske, and C. Arnosti. 2014. Composition and enzymatic function of particle-associated and free-living bacteria: a coastal/offshore comparison. ISME J. 1–13. doi:10.1038/ismej.2014.67

Elifantz, H., L. a Waidner, V. K. Michelou, M. T. Cottrell, and D. L. Kirchman. 2008. Diversity and abundance of glycosyl hydrolase family 5 in the North Atlantic Ocean. FEMS Microbiol. Ecol. **63**: 316–27. doi:10.1111/j.1574-6941.2007.00429.x

Ewart, C. S., M. K. Meyers, E. R. Wallner, D. J. McGillicuddy, and C. A. Carlson. 2008. Microbial dynamics in cyclonic and anticyclonic mode-water eddies in the northwestern Sargasso Sea. Deep. Res. Part II Top. Stud. Oceanogr. **55**: 1334–1347. doi:10.1016/j.dsr2.2008.02.013

Falkowski, P. G., P. E. Biscaye, and C. Sancetta. 1994. The lateral flux of biogenic particles from the eastern North American continental margin to the North Atlantic Ocean. Deep Sea Res. Part II Top. Stud. Oceanogr. **41**: 583–601. doi:10.1016/0967-0645(94)90036-1

Falkowski, P. G., T. Fenchel, and E. F. Delong. 2008. The microbial engines that drive Earth's biogeochemical cycles. Science **320**: 1034–9. doi:10.1126/science.1153213

Fuhrman, J. A., J. a Steele, I. Hewson, M. S. Schwalbach, M. V Brown, J. L. Green, and J. H. Brown. 2008. A latitudinal diversity gradient in planktonic marine bacteria. Proc. Natl. Acad. Sci. U. S. A. **105**: 7774–8. doi:10.1073/pnas.0803070105

Gawarkiewicz, G., R. E. Todd, W. Zhang, and others. 2018. The changing nature of shelfbreak exchange revealed by the OOI Pioneer Array. Oceanography **31**: 60–70. doi:https://doi.org/10.5670/oceanog.2018.110

Gawarkiewicz, G., R. Todd, A. Plueddemann, and M. Andres. 2012. Direct interaction between the Gulf Stream and the shelf break south of New England. Sci. Rep. **2**: 477. doi:doi:10.1038/srep00553

Hoarfrost, A. 2017. shelf1234. Github Repos. doi:10.5281/zenodo.580059

Hoarfrost, A., and C. Arnosti. 2017. Heterotrophic Extracellular Enzymatic Activities in the Atlantic Ocean Follow Patterns Across Spatial and Depth Regimes. Front. Mar. Sci. **4**: 200. doi:10.3389/fmars.2017.00200

Hoppe, H. 1983. Significance of exoenzymatic activities in the ecology of brackish water: measurements by means of methylumbelliferyl-substrates. Mar. Ecol. Prog. Ser **11**: 299–308.

Kirchman, D. L. 2001. Measuring bacterial biomass production and growth rates from leucine incorporation in natural aquatic environments. In , Methods in microbiology. Methods Microbiol. **30**: 227–237.

Ladau, J., T. J. Sharpton, M. M. Finucane, and others. 2013. Global marine bacterial diversity peaks at high latitudes in winter. ISME J. **7**: 1669–77. doi:10.1038/ismej.2013.37

Landa, M., M. T. Cottrell, D. L. Kirchman, and others. 2014. Phylogenetic and structural response of heterotrophic bacteria to dissolved organic matter of different chemical composition in a continuous culture study. Environ. Microbiol. **16**: 1668–1681. doi:10.1111/1462-2920.12242

Lomas, M. W., and N. R. Bates. 2004. Potential controls on interannual partitioning of organic carbon during the winter/spring phytoplankton bloom at the Bermuda Atlantic time-series study (BATS) site. Deep Sea Res. Part I Oceanogr. Res. Pap. **51**: 1619–1636. doi:10.1016/j.dsr.2004.06.007

Martinez-Garcia, M., D. M. Brazel, B. K. Swan, and others. 2012. Capturing single cell genomes of active polysaccharide degraders: an unexpected contribution of Verrucomicrobia. PLoS One **7**: e35314. doi:10.1371/journal.pone.0035314

McGillicuddy, D. J. 2015. Mechanisms of physical-biological-biogeochemical interaction at the oceanic mesoscale. Ann. Rev. Mar. Sci.

Monim, M. 2017. Seasonal and Inter-annual Variability of Gulf Stream Warm Core Rings from 2000 to 2016. University of Massachusetts-Dartmouth.

Mouriño-Carballido, B. 2009. Eddy-driven pulses of respiration in the Sargasso Sea. Deep. Res. Part I Oceanogr. Res. Pap. **56**: 1242–1250. doi:10.1016/j.dsr.2009.03.001

Nelson, C. E., C. A. Carlson, C. S. Ewart, and E. R. Halewood. 2014. Community differentiation and population enrichment of Sargasso Sea bacterioplankton in the euphotic zone of a mesoscale mode-water eddy. Environ. Microbiol. **16**: 871–887. doi:10.1111/1462-2920.12241

Obayashi, Y., and S. Suzuki. 2005. Proteolytic enzymes in coastal surface seawater: Significant activity of endopeptidases and exopeptidases. Limnol. Oceanogr. **50**: 722–726. doi:10.4319/lo.2005.50.2.0722

Redfield, A. C. 1958. The biological control of chemical factors in the environment. Am. Sci. 205–221.

Steen, A. D., K. Ziervogel, S. Ghobrial, and C. Arnosti. 2012. Functional variation among polysaccharide-hydrolyzing microbial communities in the Gulf of Mexico. Mar. Chem. **138–139**: 13–20. doi:10.1016/j.marchem.2012.06.001

Ullman, D. S., D. L. Codiga, A. Pfeiffer-Herbert, and C. R. Kincaid. 2014. An anomalous near-bottom cross-shelf intrusion of slope water on the southern New England continental shelf. J. Geophys. Res. - Ocean. **119**: 1739–1753. doi:doi:10.1002/2013JC009259

Wang, Y., R. Zhang, Z. He, J. D. Van Nostrand, Q. Zheng, J. Zhou, and N. Jiao. 2017. Functional gene diversity and metabolic potential of the microbial community in an estuary-shelf environment. Front. Microbiol. **8**: 1–12. doi:10.3389/fmicb.2017.01153

Wegner, C.-E., T. Richter-Heitmann, A. Klindworth, C. Klockow, M. Richter, T. Achstetter, F. Glöckner, and J. Harder. 2013. Expression of sulfatases in Rhodopirellula baltica and the diversity of sulfatases in the genus Rhodopirellula. Mar. Genomics **9**: 51–61. doi:10.1016/j.margen.2012.12.001

Zhang, W. G., and G. Gawarkiewicz. 2015. Dynamics of the direct intrusion of Gulf Stream ring water onto the Mid-Atlantic Bight shelf. Geophys. Res. Lett. **42**: 7687–7695. doi:10.1002/2015GL065530

Zinger, L., L. a Amaral-Zettler, J. A. Fuhrman, and others. 2011. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. PLoS One **6**: e24570. doi:10.1371/journal.pone.0024570

CHAPTER 3: GLOBAL ECOTYPES IN THE UBIQUITOUS MARINE SAR86 CLADE

Adrienne Hoarfrost[1], Stephen Nayfach[2], Christopher L Dupont[3], Joshua Ladau[2], Shibu Yooseph[4], C. Arnosti[1], Katherine Pollard[5]

[1] Dept. of Marine Sciences, University of North Carolina, Chapel Hill, NC

[2] Joint Genome Institute, Walnut Creek, CA

[3] J. Craig Venter Institute, Rockville, MD

[4] College of Engineering and Computer Science, University of Central Florida, Orlando, FL

[5] Gladstone Institutes, University of California San Francisco, San Francisco, CA

In prep for *ISME Journal*

Introduction

Microbial communities are important drivers of biogeochemical cycling and ecological function, and regulate the flux of carbon and nutrients in the oceans (Azam 1998; Falkowski et al. 2008). However, the complexity of interactions within microbial communities and with their environment limit our ability to link microbial community structure, function, and their variations across environments (Widder et al. 2016). Many studies have demonstrated the link between the taxonomic composition of microbial communities and their functional capabilities (e.g. Delong et al. 2006; Guidi et al. 2015; Raes et al. 2014; Shi et al. 2011; Widder et al. 2016) as well as their dependence on environmental conditions (Louca et al. 2016a; Raes et al. 2014; Sunagawa et al. 2015). However, our understanding of the geographic distributions of genetic diversity within key taxa, their relationship to environmental conditions, and the manner in which these distributions may result in functionally distinct outcomes across different environmental regions, remains limited. These limitations restrict our ability to predict genetic diversity in the environment, and in the absence of accurate models linking environmental and microbial variables, lead to ecosystem models that ignore biology entirely or make erroneous assumptions based on simple environmental relationships (Treseder et al. 2012; Wieder et al. 2015).

The biogeographical distribution of marine microbial communities has been observed at scales from single depth profiles (Delong et al. 2006) to global distributions (Ladau et al. 2013; Martiny et al. 2006), and results in spatial and temporal patterns in structure (Martiny et al. 2006; Zinger et al. 2011), function (Jiang et al. 2012; Louca et al. 2016a), and microbial diversity (Ladau et al. 2013). Mapping biogeographical distributions of microbial ecotypes from environmental variables is therefore of significant interest, illuminating the link between microbial structure, function, and ecosystem processes and enabling predictions of changes in biological phenomena as environmental conditions shift. However, there have been very few efforts to predict biogeographical patterns of genetic diversity of key microbial taxa in the ocean at large spatial scales (Kent et al. 2016; Ladau et al. 2013).

SAR86 is one such key taxa in the ocean. SAR86 is a ubiquitous marine heterotroph frequently identified in marine surface waters, and has been classified as a clade through 16S surveys of shallow ocean regions (Britschgi & Giovannoni 1991; Suzuki et al. 2001; Treusch et al. 2009). SAR86 is a very diverse group (Sunagawa et al. 2015) with at least three subclades

(Suzuki et al. 2001; Treusch et al. 2009). Despite its ubiquity in marine systems, SAR86 eludes cultivation, and knowledge of the ecological role of SAR86 in marine microbial communities is limited to genomic evidence from five genomes curated from single-cell sequencing or metagenomic assembly sources (Dupont et al. 2012; Rusch et al. 2013). Additionally, although SAR86 is very commonly detected in ocean environments, very little is known about how the distribution of subspecies and the vast genetic diversity within the SAR86 pangenome may vary across large spatial scales, and what environmental factors govern their biogeographical distribution.

In this study, we build a custom pangenome of SAR86 genes from metagenomic co-assemblies and all available reference genomes. We find that SAR86 gene presence or absence across hundreds of globally-distributed metagenomic samples are strongly associated with environmental variables. We use machine learning to build models that accurately predict the presence of SAR86 genes from environmental data from satellite and historical sources available at global resolution, and make predictions at a global scale. Machine learning, a branch of statistics that iteratively "learns" patterns from data without being explicitly programmed, is particularly well suited to this application, enabling patterns in the environmental variables that best predict SAR86 gene distributions to emerge from the global metagenomic dataset. We then identify five clusters of genes that are characterized by similar environmental distributions. Biogeographical clusters underlying the SAR86 pangenome reveal previously unrecognized ecotypes within the SAR86 clade, and a previously unappreciated geospatial complexity in this otherwise ubiquitous marine heterotroph, while patterns of taxonomic and functional enrichment across clusters hold the potential to illuminate structure-function relationships across the marine environment.

Materials & Methods

Creation of the SAR86 pangenome

A custom pangenome of SAR86 genes was created from 5 partial or near-complete reference genomes (SAR86A-E, Dupont et al. 2012; Rusch et al. 2013), 4 single-cell sequencing draft genomes (PATRIC ID 1007118.3, 1007119.3, 1009425.3, and 1009426.3; Wattam et al. 2014), as well as SAR86 genes from a massive co-assembly of metagenomes that we performed using shotgun sequencing from the Global Ocean Survey (GOS) (Rusch et al. 2007). The pangenome of 58 423 SAR86 genes was created with the MIDAS tool (Nayfach et al. 2016), resulting in a total of 51 711 nonredundant SAR86 genes clustered at 90% nucleotide identity. Only 4 188 of these genes were derived from the five reference SAR86 genomes, highlighting the diversity of SAR86 genes that are still without genomic reference.

The metagenomic co-assembly of SAR86 was generated from 226 GOS sites to create a *de novo* SAR86 database (Genbank Bioproject PRJNA13694 and European Bioinformatics Institute accession numbers ERX913362-ERX913706). All pyrosequencing and Sanger metagenomic sequences were co-assembled using the CELERA assembler (Miller et al. 2008) at 92% nucleotide identity. This threshold allowed for consensus assemblies at the species and strain level (Swan et al. 2013) with reasonable computation times. The resulting scaffolds encompass 3 Gbp of contiguous DNA sequence, while 85% of the sequence reads could be mapped back to the assembly. Open reading frames (ORFs) on scaffolds were called using MetaGene (Noguchi et al. 2006). To determine the putative phylogenetic origin of the scaffolds, each predicted peptide was phylogenetically annotated using Automated Phylogenetic Inference System (APIS, Dupont et al. 2014), which annotates according to the position of the peptide within a phylogenetic tree. Thus a peptide 99% similar to a SAR86 protein will be annotated as SAR86 (with the associated taxonomic tree), while a peptide that branches basally within the phylogenetic tree next to *Gammaproteobacteria* would only be annotated as such. The scaffolds were taxonomically annotated at the lowest level for which greater than 50% of the ORFs had agreement in the APIS calls. This approach has been used in previously published pangenome biogeographic analysis (Kent et al. 2016).

Mapping SAR86 gene presence/absence in a global metagenomic dataset

Sequencing reads from the TARA datasets were mapped to the SAR86 pangenome to determine SAR86 gene presence/absence at each TARA site. The TARA project (Sunagawa et al. 2015) was a large-scale, multi-year effort to sequence hundreds of marine seawater samples from globally-distributed sites and from depths ranging from surface to mesopelagic using consistent techniques (SI Fig 3.1). To date, the TARA project has made available 243 individual metagenomic samples, 198 of which were for live samples (pore size >0.22) and 45 of which were <0.22 filter pore size blank controls. The 198 live TARA samples were used to map to the SAR86 pangenome and create a global dataset of SAR86 gene presence/absence. Specifically, MIDAS (v1.1.0, Nayfach et al. 2016) was used to perform read alignment with bowtie2 (options --very-sensitive), mapping reads according to their best-hit with a % DNA identity of >90%. Reads with mapping quality <20 or mean base quality <20 were discarded. Where multiple sequencing runs had been conducted for a single sample, the run with the highest number of sequencing reads was used.

This mapping procedure resulted in gene coverage values for each of the 51 711 SAR86 genes at each of the 198 live TARA sites. Gene coverage was normalized to a mean coverage value of 1 for each site by dividing by the mean gene coverage at that site, and finally, this continuous value was converted to binary gene presence/absence at a threshold of 0.37, where genes with a normalized coverage value less than or equal to 0.37 were considered absent from the sample, and those with coverage greater than 0.37 were considered present. This threshold value was chosen from manual inspection of the SAR86 gene coverage distribution at individual TARA sites; where a strain or mixture of strains of SAR86 are present, a peak in normalized gene coverage frequency is observable, while erroneous gene assignments form a long tail of low coverage. The gene coverage values of this erroneous long tail was typically below e^{-1}, or 0.37, as was particularly visible when viewed as a natural log distribution of gene coverage, so this value was chosen as a threshold below which normalized SAR86 genes are ‘absent’.

Remote environmental data curation and processing

In order to build models predicting SAR86 gene presence from environmental data on a global scale, environmental features available at global resolution were collected from a combination of contemporary satellite data and curated sources of historical averages. A total of 51 features were collected. These features, their sources, definitions, and units are summarized in SI Table 3.1. These data included 6 features from contemporary satellite data – sea surface temperature, chlorophyll *a* concentration, photosynthetically active radiation at the sea surface, particulate inorganic carbon at the sea surface, particulate organic carbon at the sea surface, and net primary productivity at the sea surface. 3 features constituted data about the source of the sample: latitude, longitude, and depth sampled. The remaining features were environmental data available at global resolution from historical averages of many values of interest, such as day length, dust flux, pycnocline depth, apparent oxygen utilization and concentrations of nitrate, phosphate, and silicate (SI Table 3.1). Historical environmental data were typically an average value at the annual, monthly, or decadal scale, averaged over a historical period of time ranging from decadal to 57-year time scales. For example, the historical annual mean pH value, derived from BioOracle (Tyberghein et al. 2012), is the average annual pH over the period 1955-2012; the historical monthly mean nitrogen:phosphate ratio, in contrast, is the average N:P ratio during a particular month, averaged over the period of 1955-2012. The annual standard deviation for a historical value, available for some historical environmental features, gives an idea of the typical variation in an environmental variable over the course of a year for the relevant historical period.

For each TARA site, the environmental feature value closest to the sampling site's latitude, longitude, and, where relevant, the sampling depth and/or sampling date, was derived from the original environmental data source. Environmental features, which are all continuous variables, were then centered and normalized to a mean of zero and a standard deviation of one across all TARA sites. This preprocessed environmental feature matrix (SI Table 3.2) served as the input feature vectors for each TARA site during model training.

Gene presence/absence models & predictions

To predict SAR86 gene presence or absence from environmental variables, classification models were built using logistic regression with L1 regularization for each SAR86 gene, using the remote environmental data across TARA sites as input features. First, TARA sites where SAR86 was not present or abundance was low were filtered out, defined as sites where fewer than 1000 genes had 5x coverage (22 out of 198 TARA sites). Sites where environmental data features were missing were also filtered out (an additional 21 out of 198 TARA sites). This resulted in 155 out of the original 198 TARA sites that were used to train the models. 20 out of 22 of TARA sites where SAR86 was not present or was in low abundance were mesopelagic samples; however, 41 TARA sites from mesopelagic depths had SAR86 abundances high enough to be included in model training. This set of 155 TARA sites was randomly split into a training, validation, and test sets of 111, 13, and 31 sites respectively.

Of the 51 711 genes in the SAR86 pangenome, 24 317 of these were present at 20-80% of TARA sites. Logistic regression models were trained only for these variable genes, since it was not meaningful to predict the distribution of very rare or very common genes.

The L1 regularization penalty was tuned to minimize overfitting while maximizing accuracy in the validation set. A penalty parameter of 0.7 was chosen, which achieved a mean validation accuracy of 82.1% across all models, with 66.9% of models overfit. This is in comparison to no regularization penalty (C=1.0), which achieves 82.2% accuracy in the validation set and overfits in 70.1% of models.

A final logistic regression model was trained independently for all 24 317 variable genes, with a L1 regularization penalty (C=0.7), using the scikit-learn python package. These models can be reproduced with code available on the associated Github repository (Hoarfrost 2018). The recall, or the sensitivity rate, is calculated as (true+/(true+ + false-); precision is defined as (true+/(true+ + false+); and the F1-score is defined as 2*((precision*recall)/(precision+recall)).

Clustering

The coefficients associated with each environmental feature input to the logistic regression models were used to cluster SAR86 genes into groups that were best predicted by similar environmental variables. By clustering on environmental variables associated with gene models rather than, for example, gene covariation across TARA sites, environmental variables underlying geographic distributions of genes are identified by definition, and projecting cluster distributions beyond TARA sites at global scales from environmental data is possible A k-means clustering algorithm was used for clustering genes into five clusters. The number of clusters, *k*, was chosen for the *k* at which the inertia (sum of the squared distance of each point to its centroid) begins to decrease less rapidly; where the distance between centroids begins to increase less rapidly; and where map projections and comparisons of Euclidean distances of cluster centroids look unique for each cluster without producing repetitive distributions. Exploratory analysis in a jupyter notebook and a python script for reproducing clusters are available on the Github repository (Hoarfrost 2018).

Global projections of the biogeographical cluster distributions used global resolution spatial data of environmental features in netCDF files to make predictions at global scale. Prediction values were derived by multiplying normalized environmental feature values for each latitude and longitude coordinate by the cluster centroid coefficient for that environmental feature, summing this vector of coefficient-multiplied environmental feature values, and converting to a [0,1] scale using a sigmoid function, where any value greater than or equal to 0.5 is considered a ‘present’ prediction, and anything less than 0.5 is considered ‘absent’ (Hoarfrost 2018).

Taxonomic & functional enrichment analysis

The distribution and enrichment across clusters were evaluated at the genome, contig, and functional level for two SAR86 reference genomes, SAR86A and SAR86E, for the contigs of the SAR86 co-assembly, and for the functional annotations to Pfam (Finn et al. 2016) for the SAR86 pangenome.

Genome distribution across clusters was evaluated by counting the percentage of genes in the SAR86 pangenome originating from SAR86A or SAR86E assigned to each cluster. The correlation of SAR86A/E relative abundance with their associated cluster proportion within TARA sites was evaluated by deriving the relative abundance of SAR86A and E from the mapped TARA samples, and renormalizing such that the abundances of SAR86A and E sum to 1. The cluster proportions of cluster 1+5 and cluster 3+4 were similarly renormalized to sum to 1.

Enrichment values were calculated the same way for both contig and functional enrichment analyses. The enrichment of a particular contig or Pfam family annotation was calculated as the actual number of genes from a contig/annotation observed for a cluster minus the expected value for a cluster, all divided by the expected value for that cluster. For example, for a particular contig, the expected value for a particular cluster is the number of genes on that contig proportionally distributed across clusters – or, on cluster 1, is the number of genes on the contig multiplied by the percentage of total SAR86 genes assigned to cluster 1. The actual number of genes from a contig for cluster 1 is simply the number of genes from that contig assigned to cluster 1. So, the final enrichment value is (actual – expected)/expected. A contig or Pfam was considered enriched on a cluster if the number of genes assigned to that cluster exceeded the value expected if genes from a contig were assigned evenly across clusters. An enrichment value above zero is therefore enriched, with a value of e.g. 2.1 resulting where 210% more genes from that contig were assigned to a cluster than expected; whereas a value below zero is depleted relative to the expected value, and a minimum value of -1 is found where zero genes were assigned to a cluster. An enrichment value of 0 indicates that the contig or annotation is not enriched or depleted for that cluster and is identical to what would be expected if genes were randomly assigned to clusters.

In the case of the functional enrichment analysis, there were 1337 Pfam families to which at least one gene was annotated; however, only 405 Pfams for which more than 20 genes were annotated to it were used for the enrichment analysis, since a rare Pfam with only one gene annotated to it will look perfectly enriched to whichever cluster that gene was assigned to and skew enrichment analysis results.

To test the statistical significance of cluster enrichment values, a nonparametric Mann-Whitney U test was applied to test whether the vector of enrichment values associated with each contig/annotation for each cluster was significantly different from the expected distribution of all-zeros.

Results

This study first built machine learning models to learn the relationships between SAR86 gene distributions and environmental variables. The regularized logistic regression approach used enabled us to identify the subset of environmental variables that are most important for predicting the geographical distributions of each individual gene, and the coefficient associated with that environmental variable-gene relationship. Using unsupervised clustering, we then identify clusters of genes with similar environmental distributions. Clustering is an approach that enables us to identify emergent properties and structure underlying the environmental gene distributions without explicit prior knowledge of expected SAR86 ecotypes. By using environmental variables available at global resolution in the original gene distribution models, we are then able to forecast these emergent properties at spatial scales far beyond the sampling locations specific to this study.

Accurate prediction of SAR86 gene distributions from environmental variables

SAR86 gene content in TARA Oceans metagenomes is associated with environmental characteristics of sampling locations. We built a regularized logistic regression model for each gene that captures the probability of the gene being present in the SAR86 genomes at a given location as a function of environmental variables. The L1 regularization we used during model training selects the most predictive environmental features, while the coefficients for less predictive environmental features converge to zero, effectively ignoring these features. This enables us to identify which of the many environmental variables are most reliably associated with a gene's presence while avoiding overfitting.

The resulting 24 317 gene models predicted SAR86 gene presence/absence with an average of 79.4% accuracy in the test set, and a median test accuracy of 80.6%. From the confusion matrix (Fig 3.1a), we see that precision and recall measures are roughly even (0.85 and 0.81, respectively), with an F1 score of 0.83. 87.4% of models, for 21 264 out of 24 317 gene models, had accuracies in the test set that were an improvement over the majority class accuracy, or the accuracy of the model if it predicts 'all absent' or 'all present', whichever is in the majority (Fig 3.1b).

As an additional test of the robustness of the models, the accuracy of predictions at those TARA sites that were not included in model development, where SAR86 was not present or were in very low abundance, was also examined. There were 20 of these TARA sites for which environmental data was available for all features. At these sites, 87.6% of the 24 317 SAR86 genes were actually absent, while the gene models predicted that 65.2% of genes were absent. Overall, the average accuracy across the gene models was 68.5%, while the median accuracy was 70.0%. This is less accurate than the performance of the models at sites where SAR86 was present, but still reasonably accurate, and suggests that these models are robust to predictions outside of the distribution of gene presence used in training the models.

An average of 17 of 51 environmental features was significantly associated with each gene's distribution across TARA Oceans sites, with many features shared across genes (i.e., frequently selected during model training for both genes) (SI Fig 3.2). These include latitude, longitude, distance from land, ocean depth, and other features that might describe the general ocean basin or region of a sample; as well as pH, sea surface temperature, pycnocline depth, nitrogen:phosphorous ratio, cloud fraction, or other environmental factors that describe regions of the ocean that experience particular conditions at more fine resolution.

While the environmental features that best predict gene presence/absence varies by the individual gene model, and many of the 51 environmental variables are covariates with one another, training logistic regression multiple times on the same data results in the same sets of environmental features being chosen as the most predictive for each gene model (see jupyter notebook in Hoarfrost 2018). This consistency suggests that the environmental features selected in each model reflects a true difference in predictive power between the selected features and those that were not selected, rather than a random choice among features that are roughly equally predictive.

Clustering of SAR86 genes into common environmental distributions & global projections of their biogeographic distributions

The environmental features that best predict individual genes, and the strength of the coefficients associated with any particular environmental feature, vary by the individual gene model. However, there are apparent patterns among genes as to which environmental variables are most predictive, and the magnitude and sign of the coefficients associated with those variables, with some groups of genes appearing to be predicted by similar environmental features. This suggests that genes which are predicted by similar environmental features in similar ways occupy similar geographical distributions characterized by unique environmental conditions.

Kmeans clustering of genes by their logistic regression environmental feature coefficients revealed five clusters within the SAR86 pangenome characterized by similar environmental distributions. Each cluster reveals a distinct set of environmental features and coefficient patterns associated with the centroid of that cluster (Fig 3.2, SI Table 3.3). The centroids of the clusters define the average coefficient associated with each environmental feature across the genes that make up that cluster (SI Table 3.3). The cluster centroids can be used to project cluster gene presence or absence at a global scale, using the known global distributions of the input environmental features (Fig 3.3). These global projections of SAR86 biogeographical clusters reveal differential patterns in predicted geospatial distributions of SAR86 genes that are characteristic of the environmental features that best predict the presence of genes in that cluster.

In cluster 1, the environmental features with the highest magnitude coefficients included oceantemp_monthly_historical (-1.23), dustflux_monthly_historical (-0.75), phosphate_annual_historical (-0.66), pycnoclinedepth_monthly_historical (0.50), and silicate_annual_historical (-0.40). The high magnitude of monthly historical features suggests a seasonal component, with genes more likely to be present during the winter and spring, while the negative coefficients associated with nutrients, temperature, and dust flux confine projections primarily to open ocean temperature regions.

In cluster 2, the environmental features with the highest magnitude coefficients included longitude (-1.01), par_annual_historical (0.84), cloudfraction_monthly_historical (-0.72), and solarinsolation_annualstdev_historical (-0.40). The photosynthetically active radiation (PAR), cloud fraction, and solar insolation coefficients all select for higher likelihood of gene presence in areas of high light, few clouds year round, and low variability in sunlight, as is typically found in lower latitudes, while the longitude coefficient selects for locations in the western hemisphere.

The TARA samples from the Pacific Ocean were only sampled from the western hemisphere, and were also the only locations where mesopelagic samples were taken. Not coincidentally, the TARA sites with the highest proportion of cluster 2 genes present were the mesopelagic samples and those samples taken in the Pacific Ocean (Fig 3.4).

Cluster 3's most predictive environmental features included diffuseattenuation_annual_historical (-2.43), longitude (0.92), sst_annual_historical (0.87), and dustflux_monthly_historical (-0.40). The global projections for cluster 3 predict gene presence most confidently in the eastern hemisphere, at lower latitudes where temperature is higher, and away from coasts where diffuse attenuation is lower.

Cluster 4 environmental features with highest magnitude coefficients included sst_annual_historical (0.95), dustflux_monthly_historical (-0.53), longitude (0.42), ph_annual_historical (-0.41), and thermoclinedepth_annualstdev_historical (-0.37). Similarly to cluster 3, the genes in cluster 4 are predicted more likely to be present in the eastern hemisphere and where temperature is higher. There is also a slightly higher likelihood of a gene being present in the southern hemisphere, where dust flux is lower. The coefficients of features associated with cluster 4 are more even than other clusters, and so additional features with lower coefficient magnitudes, such as a negative association with phosphate (phosphate_monthly_historical , -0.26) and a positive association with cloud fraction variation (cloudfraction_annualstdev_historical, 0.22) also help to explain the predicted distribution of cluster 4 genes.

Cluster 5 environmental features with highest magnitude coefficients included solarinsolation_annualstdev_historical (0.58), sst_annual_historical (-0.49), pic_satellite (-0.42), cloudfraction_annual_historical (0.41), and silicate_annual_historical (-0.36). The positive association with variable sunlight (solarinsolation_annualstdev_historical) and negative association with temperature result in distributions of genes present at higher latitudes, which is also contributed to by the positive association with cloud fraction, which is generally concentrated at higher latitudes and at the equator. The negative association with silicate restrict the southern polar distribution such that genes are less likely to be present near the Antarctic Convergence/AA Polar Front.

Each TARA site contained genes from a mixture of clusters, but the dominant clusters and the evenness of the proportion of each cluster was variable across sites (Fig 3.4, SI Fig 3.3, SI Table 3.4). For example, the predicted geographic distribution of cluster 2 as higher in the western hemisphere and Pacific Ocean is evident in the cluster proportions across TARA samples, for which cluster 2 is present in highest proportions for those TARA sites sampled in the Pacific Ocean (Fig 3.4, SI Fig 3.3b). In contrast, cluster 3 genes are found in higher proportions at TARA sites sampled in the Eastern hemisphere, reflecting their predicted geographic distributions (Fig 3.4, SI Fig 3.3c). The Shannon diversity measuring the relative proportions of the five clusters at TARA sites, which accounts for both the relative evenness of the cluster proportions as well as how many of the five clusters are present, ranged from 0.699 to 1.532 (SI Table 3.4). The minimum possible value of 0 would indicate a site at which genes from only one cluster was present, while the maximum possible value of 1.609 (*ln(5)*) would indicate a site where genes from all five clusters are present in equal proportions. The TARA sites with the lowest Shannon diversity metrics include TARA station 93 at 34°S and 73°W off the coast of Chile, which is dominated by cluster 5 genes, and TARA stations 38, 42, 45, and 36 in the Indian Ocean, which are dominated by cluster 4 genes. The TARA sites with the highest Shannon diversity metrics include many of the mesopelagic depth samples in the Pacific Ocean, as well as all depths sampled at station 70 in the South Atlantic basin at 20.4°S and 3.2°W.

Taxonomic enrichment & functional differentiation across clusters define SAR86 ecotypes

The cluster assignments of genes from the SAR86 references genomes showed clear partitioning on taxonomic lines. Of the 24 317 genes that were used in models, only 810 originated from one of the five reference SAR86 genomes. Of these, 622 genes were from SAR86A, and 157 genes originated from SAR86E, while only 5, 7, and 19 genes originated from SAR86C, D, and B respectively. The cluster assignments of genes from SAR86A and SAR86E were clearly differentiated, with genes from each genome assigned primarily to two clusters, and each cluster dominated by one genome. SAR86A genes were partitioned primarily into clusters 4 and 3, with 493 and 118 out of the 622 SAR86A genes assigned to cluster 4 and 3 respectively, while only 4 and 7 genes were assigned to clusters 2 and 5, and 0 genes to cluster 1. The 157 SAR86E genes were partitioned into clusters 1 and 5, with 76 and 78 genes respectively, while only 2 and 1 genes were assigned to clusters 2 and 4, and 0 genes to cluster 3.

Clusters also showed clear taxonomic differentiation at the contig level. Those genes that did not originate from one of the five SAR86 genomes originated from one of 732 contigs from the SAR86 co-assembly. Genes from the same contig were generally assigned to the same cluster, such that gene assignments of almost all contigs, 540 out of 732 contigs, were enriched on only one cluster, 183 contigs were enriched on only two clusters, and the remaining 9 contigs were enriched on 3 clusters (Fig 3.5). Where a contig was enriched, the enrichment was strong, with an average enrichment of 3.03 and a standard deviation of 0.43, and ranging from 1.41 in cluster 4 to 5.25 in cluster 2. This strong enrichment for one or a subset of clusters was paired with a high frequency of -1.0 depletion values on the other clusters, which indicates that no genes from a contig were assigned to that particular cluster: 522 out of 732 contigs had -1.0 enrichment values in cluster 1, 560 in cluster 2, 561 in cluster 3, 350 in cluster 4, and 279 in cluster 5. A Mann-Whitney U test testing whether the contig enrichment values for each cluster were significantly different than the null enrichment distribution of all zeros was highly significant for all clusters, with p values <0.001 for all clusters and as low as 7.4×10^{-159} for cluster 2 (Fig 3.5c).

The taxonomic partitioning across clusters is further supported by the relationship between cluster proportions and the relative abundances of SAR86 genomes at TARA sites. The clusters associated with SAR86A, clusters 3 and 4, were in higher proportions relative to the clusters associated with SAR86E, clusters 1 and 5, at TARA sites where SAR86A abundances were higher relative to SAR86E (SI Fig 3.4, Pearson $R^2 = 0.70$, $P=1.56 \times 10^{-26}$).

In addition to taxonomic enrichment across clusters, there was also significant partitioning of genes at the functional level, with differential enrichment of Pfam annotated genes across clusters (Fig 3.6). Pfams were enriched on average by an enrichment value of 0.25 and a standard deviation of 0.10, ranging from 0.13 in cluster 4 to 0.32 in cluster 2. This enrichment was significant for most of the clusters, with a Mann-Whitney U test resulting in p values of 5.6×10^{-3} for cluster 1, 1.6×10^{-9} for cluster 2, 0.016 for cluster 3, 1.1×10^{-11} for cluster 4, and 0.11 for cluster 5 (Fig 3.5c). Using a Bonferroni-corrected p value cutoff of 0.01 for significance, this suggests that clusters 1, 2, and 4 have significant functional enrichment, while functional enrichment on cluster 3 is marginally significant. Genes from a particular Pfam were most often assigned to only two or three clusters (Fig 3.6b). While functional enrichment in general was less strong than taxonomic enrichment, this may be due to the relative coarseness of functional annotation compared to taxonomic assignments, and our inability to annotate many genes with confidence. Enrichment of specific Pfams corresponding to some ecologically important heterotrophic functions indicated possible niche differentiation across clusters: Glycosyl hydrolase family 3, which corresponds to exo-acting glucosidases, was enriched across clusters 3, 4, and 5, and depleted in clusters 1 and 2, while glycosyl hydrolase family 16, which corresponds to endo-acting glucanases, was enriched strongly on cluster 3, depleted in clusters 1 and 2, and near the null value for clusters 4 and 5 (SI Fig 3.5).

Discussion

While SAR86 is generally considered to be a ubiquitous heterotroph, this study demonstrates that SAR86 genetic diversity is strongly associated with, and accurately predicted by, environmental variables, and that distinct environmental distributions of gene clusters define a deeper geographic variability of SAR86 subgroups than previously appreciated. The SAR86 clade is a group within the *Gammaproteobacteria* classified as such by their 16S rRNA gene similarity (Britschgi et al. 1991; Suzuki et al. 2001; Treusch et al. 2009). The three near-complete and two partial genomes available for SAR86 (Dupont et al. 2012; Rusch et al. 2013) show high diversity within this clade; average nucleotide identity between genomes is between 70-80% (SI Table 3.5). In light of this high diversity, it is not surprising that the SAR86 pangenome can be disentangled into five distinct clusters with different geographic distributions associated with unique environmental predictors. These clusters are differentiated at the taxonomic and functional level, which has implications for our interpretation of SAR86, its biogeographic distribution, and its ecological role within microbial communities and across the marine environment.

Using a data intensive approach to build machine learning models of the relationship between SAR86 genes and environmental variables at a global scale, we demonstrate how such an approach can be used to better understand the dynamics driving microbial biogeography, revealing patterns that may have been missed at the 16S OTU or community level, or with data from a smaller scale. Particularly as microbial data becomes increasingly available in the future, such an approach holds promise for illuminating the relationship between microbial community structure and ecological function across broad spatial scales.

An ecotype (Cohan 2006) is often identified in practice as a group of closely related lineages that co-occur on the same spatial or temporal scale and are associated with particular environmental conditions. The results of this study identify clusters of genes that, while their phylogenetic relatedness is unknown, are taxonomically and functionally differentiated and occupy distinct environmental distributions. While the functional traits that may confer niche specificity within these distributions is not obvious from our results, functional differentiation across clusters of glycosyl hydrolases (SI Fig 3.5), an important class of enzymes for heterotrophic metabolism of polysaccharides, suggest that genes associated with different clusters occupy distinct functional niches. Glycosyl hydrolase families 3 and 16 target many of the same substrates – β-linked glucans, including the abundant marine plankton storage glucan laminarin – but using different enzymatic mechanisms (Lombard et al. 2013). The strong enrichment in cluster 3, and strong depletion in clusters 1 and 2, of both families, compared to the enrichment of only family 16 in clusters 4 and 5, may indicate differing metabolic strategies of SAR86 taxa enriched on each cluster that describe distinct ecological functions. Given the clear taxonomic and functional partitioning of the SAR86 pangenome across clusters with distinct geographic distributions associated with unique environmental variables, and likely niche partitioning across clusters, we conclude that the clusters described here define previously unidentified ecotypes within the SAR86 clade.

Previous investigations of temporal and geographic patterns in SAR86 noted that while the phylogenetic substructure of the SAR86 clade implies that it may be made up of multiple ecotypes, they were not able to identify these at the limited geographic resolution of their study (Treusch et al. 2009). The potential existence of SAR86 ecotypes was also noted in the apparent biogeographic distributions of SAR86A, B, C, and D genomes (Dupont et al. 2012), which differed in their distributions across coastal vs. open ocean sampling sites and along temperature gradients. This general observation is supported by the predicted distributions of the clusters identified in our study, for which three clusters (clusters 2, 3, and 4) are partially defined by their warmer, open ocean distributions, and two (clusters 1 and 5) were associated with cooler temperatures. The difficulty of identifying ecotypes in SAR86 is in contrast to SAR11, for which distinct ecotypes have been identified within a constrained geographic sample because they were strongly associated with differences in depth and salinity distributions (Carlson et al. 2009). This study was able to identify SAR86 ecotypes, despite their partially sympatric distributions that causes a single sampling site to be composed of genes from multiple clusters, because of the larger data size and geographic distribution of the TARA dataset.

The taxonomic and functional differentiation of genes across SAR86 ecotype clusters is significant in the context of interactions between microbial community structure, function, and ecology. Both community composition (Ladau et al. 2013; Martiny et al. 2006; Pommier et al. 2007; Zinger et al. 2011) and functional traits (Delong et al. 2006; Jiang et al. 2012; Louca et al. 2016a; Shi et al. 2011) vary biogeographically and can be predicted to some extent by environmental variables (Ladau et al. 2013; Louca et al. 2016a). Taxonomic variation can lead to functional differentiation of microbial communities (Delong et al. 2006; Galand et al. 2018; Strickland et al. 2009), which ultimately shapes the biogeochemical and ecological function of a community; conversely, functional redundancy across microbial taxa can complicate the relationship between structure and function (Louca et al. 2018), with taxonomically variable communities playing similar functional roles (Louca et al. 2016b). Disentangling the relationship between environment, biogeography, structure, and function is therefore a significant ongoing challenge in microbial ecology (Louca et al. 2016a; Morales et al. 2011; Raes et al. 2014; Widder et al. 2016). By focusing on patterns at the individual gene level within a single clade, we are able to uncover patterns in environmental distributions of genetic diversity at a scale that would normally be obscured by the complexity inherent to microbial communities. For example, previous studies have found that functional classifications of taxa are better predicted by environmental parameters than taxonomic 16S-based classifications (Louca et al. 2016a); however, these functional classifications are broad – all of the SAR86 pangenome would be classified as 'aerobic chemoheterotroph' – in order to control for the inherent noise and complexity of mixed microbial communities. It is likely that within the SAR86 pangenome there is ecological niche differentiation within this category that, for example, could lead closely related phylotypes of SAR86 that occupy different ecotypes to utilize different substrates (Aguilar et al. 2004; Hunt et al. 2008; Martiny et al. 2015), which is supported by the functional enrichment across of our clusters and the differential enrichment of carbohydrate utilizing enzymes (SI Fig 3.5). Previous analyses of the genomic context of SAR86 genomes (Dupont et al. 2012) also suggest that much of the diversity between SAR86 genomes may be driven by fine scale diversification of catabolic enzymes on loci associated with TonB dependent receptors, which are responsible for transporting carbon compounds (as well as metals) into the cell.

The accuracies of our gene models are better on average than previous studies (0.79 vs 0.48, (Louca et al. 2016a), which may similarly be due in part to our focus on modeling individual genes rather than whole communities, for which there may be less noise and variation inherent to the data. This difference in model accuracy may also be due to our consideration of different, and a larger number, of input environmental features, and the fact that environmental features were chosen for their availability at global resolution rather than their putative importance in regulating microbial function. These environmental features may be more predictive of the distributions of SAR86 genes, even if they are less relevant to biological function. The environmental factors that influence whether an organism grows in a particular location or community may be different from those that drive their function within that community: for example, an organism may only grow in fresh or saline waters, while the maintenance of a nitrogen fixation pathway depends on nutrients or other factors. It is important to note that those environmental features that are selected as most predictive for each gene model do not necessarily drive the growth of SAR86 in a causal manner, but implies only that these environmental features are good predictive proxies for the presence of that gene. The interpretation of the most predictive environmental features may vary depending on the feature; some features may be a proxy for biological phenomena, while others simply define oceanographic regions, or are proxies for other factors not able to be measured that are true causal drivers of variation. The features chosen by the L1 regularization procedure are also likely biased by the scope of the samples used as inputs to the model: for example in this study, the cluster associated with western hemisphere longitudes is overrepresented in sites from the TARA expeditions in the Pacific Ocean; however, there are longitudes both east and west of the antemeridian in the Pacific, represented as negative and positive longitudes in the models, and it is a limitation of the TARA dataset that only samples from the eastern part of the basin, in the western hemisphere, are represented. This observation also serves as a note of caution for the interpretation of the global projections, whose predicted distributions will likely break down most where representation of samples is most sparse, e.g. in polar regions.

We are able to make accurate predictions of geographic distributions of SAR86 genes at a global scale, identifying previously unacknowledged biogeographical complexity within an otherwise ubiquitous heterotrophic clade and making global projections of the distributions of SAR86 ecotypes associated with distinct environmental distributions. The approach used leverages a large dataset across broad geospatial distribution, demonstrating the potential of machine learning and the use of broader scale integrated datasets for marine microbial ecology. Such an approach may also be useful for bioprospecting: for example, to identify locations at which a cultured representative for a predicted gene of interest is likely to be found. The five global ecotypes underlying the highly diverse SAR86 clade, the taxonomic and functional differentiation across ecotypes, and the distinct environmental distributions of SAR86 genetic diversity highlights the importance of SAR86 within marine microbial communities and broadens the ecological context and interpretation of the ubiquitous marine heterotroph SAR86 across the world’s oceans.

Acknowledgements

This work was supported by a Deep Carbon Observatory Deep Life Modeling & Visualization Fellowship to AH; OCE-1736772 to CA; and Gordon and Betty Moore Foundation grant #3300 to KSP.

Conflict of Interest

The authors declare no conflict of interest.

Author Contributions

CD and SY created the SAR86 co-assembly of SAR86 genes from the Global Ocean Survey sequences, and CD annotated the SAR86 pangenome. SN created the pangenome and mapped TARA samples to the SAR86 pangenome. AH gathered satellite environmental data, created the models, did clustering, identified ecotypes and analyzed data. JL gathered historical environmental data. All authors contributed to discussion of data and writing of the manuscript.

Figures

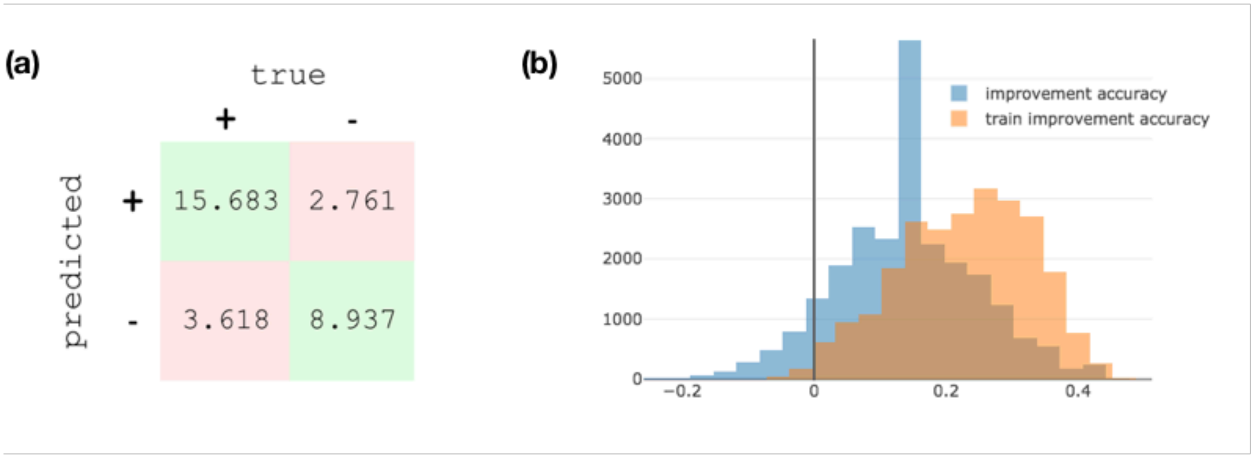


Fig. 3.1 – Average confusion matrix (a) over all gene models and histogram of improvement accuracy (b) for the test set (blue) and training set (orange). In (a), true labels are indicated by columns, while labels predicted by the models are indicated by rows. From top left to bottom right rowwise, this corresponds to true positive, false positive, false negative, and true negative predictions. In (b), improvement accuracy is defined as improvement over the majority class

accuracy. An improvement accuracy great than zero, indicated by the vertical line, corresponds with models that score better accuracies than the majority class accuracy.

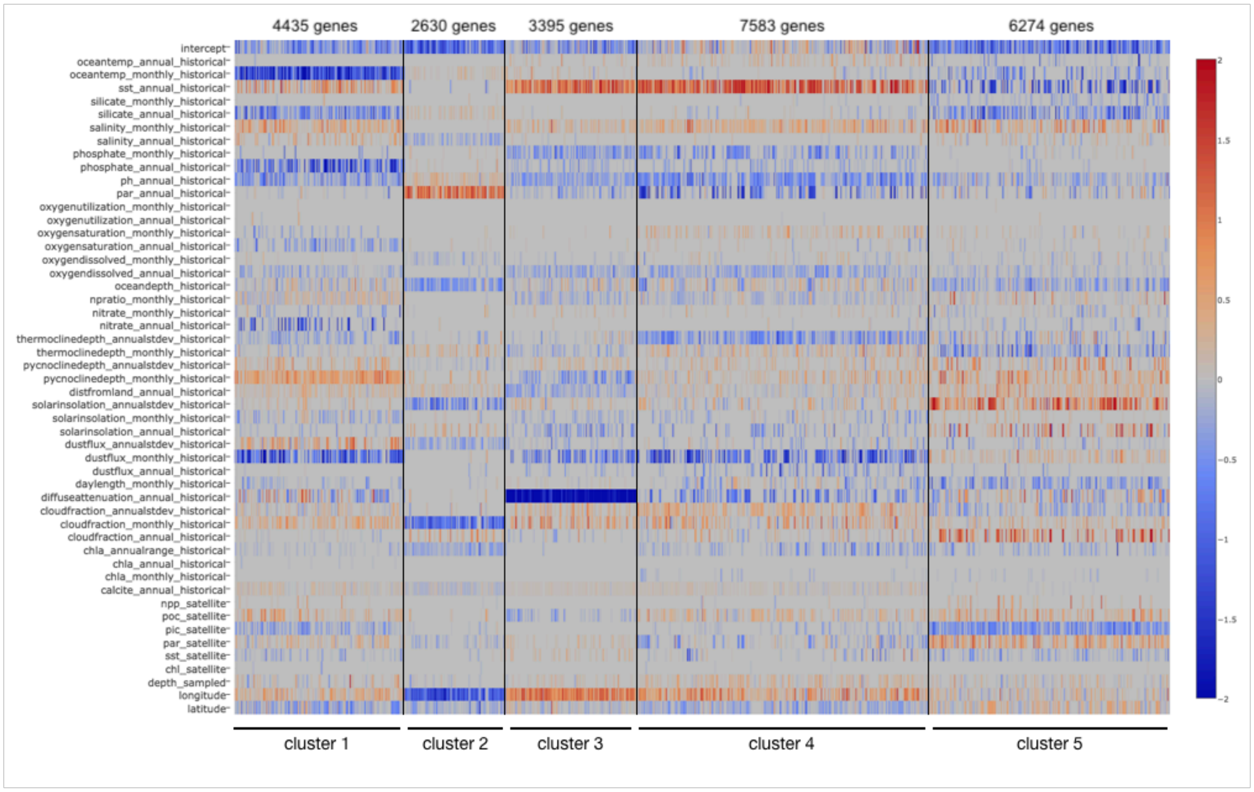


Fig. 3.2 – Heatmap of model coefficients for each environmental feature (rows) and gene (columns). Genes are ordered by their biogeographical cluster assignments (x axis).

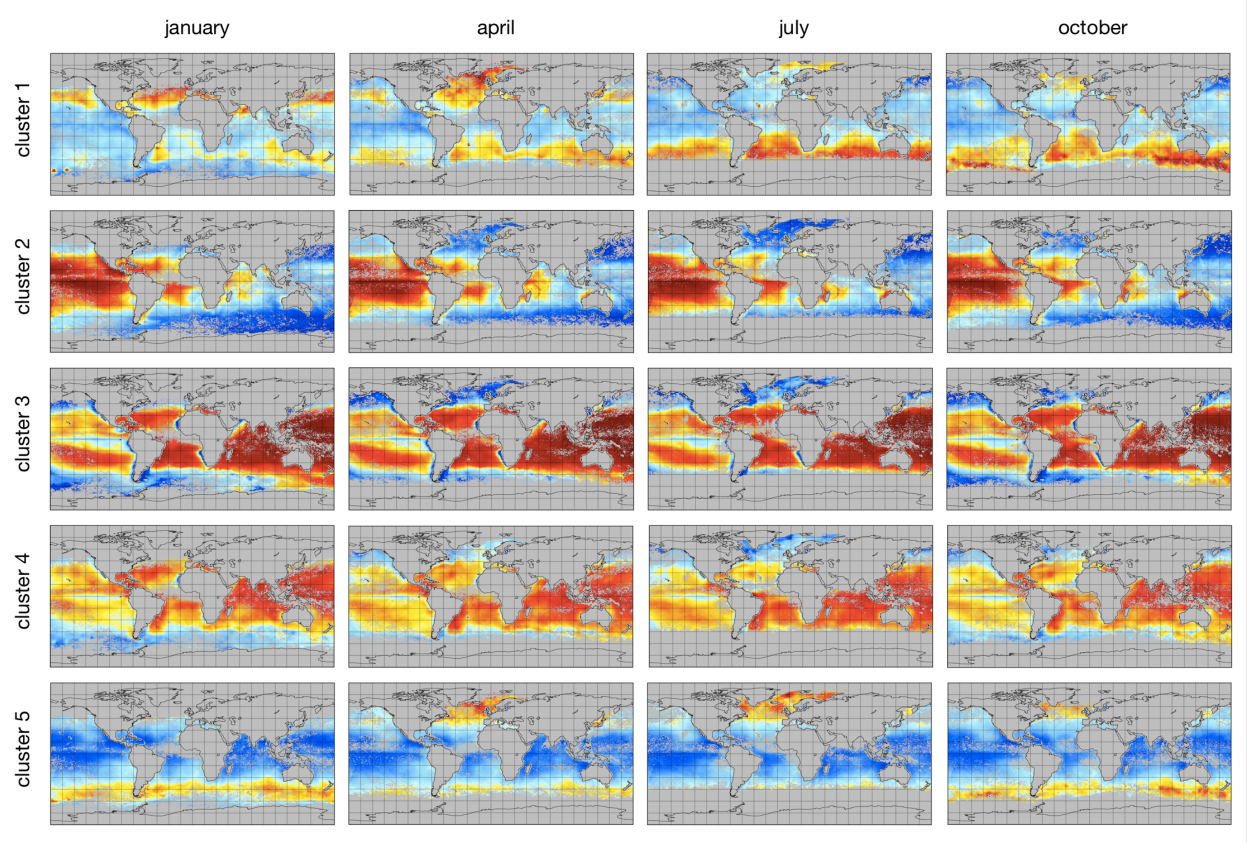


Fig. 3.3 – Global map projections of cluster distributions for each cluster (rows) in January, April, July, and October of 2009 (columns). Red indicates a high confidence of a gene cluster being present (prediction near 1), blue a high confidence of a gene cluster being absent (prediction near 0), and white a low confidence prediction (prediction of 0.5).

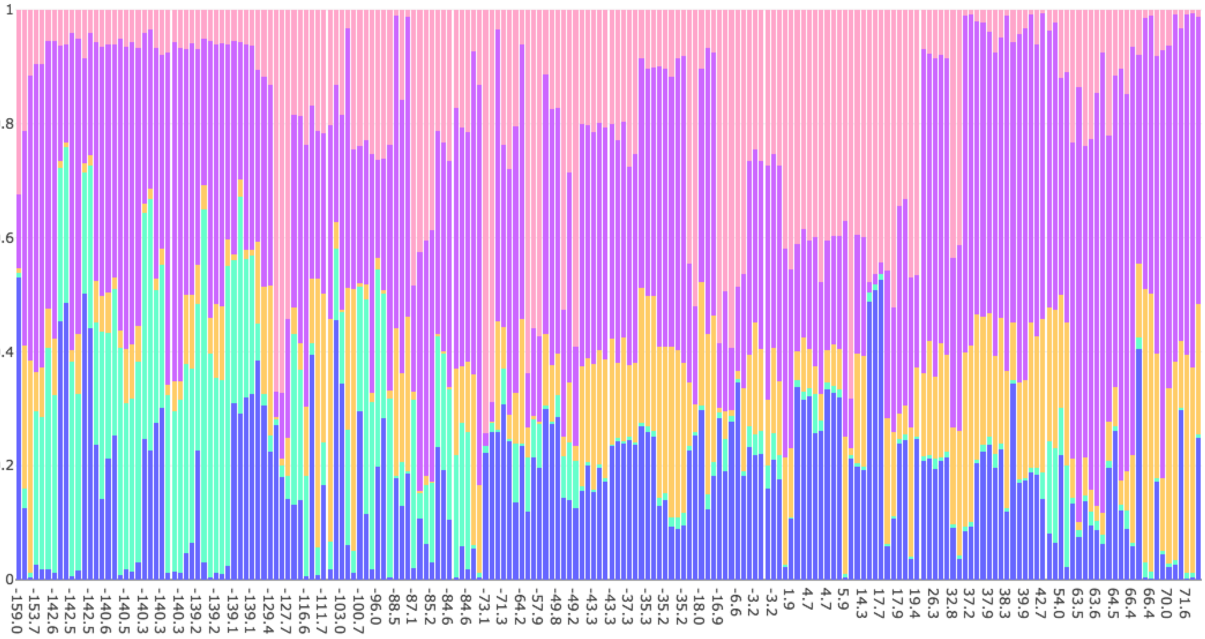
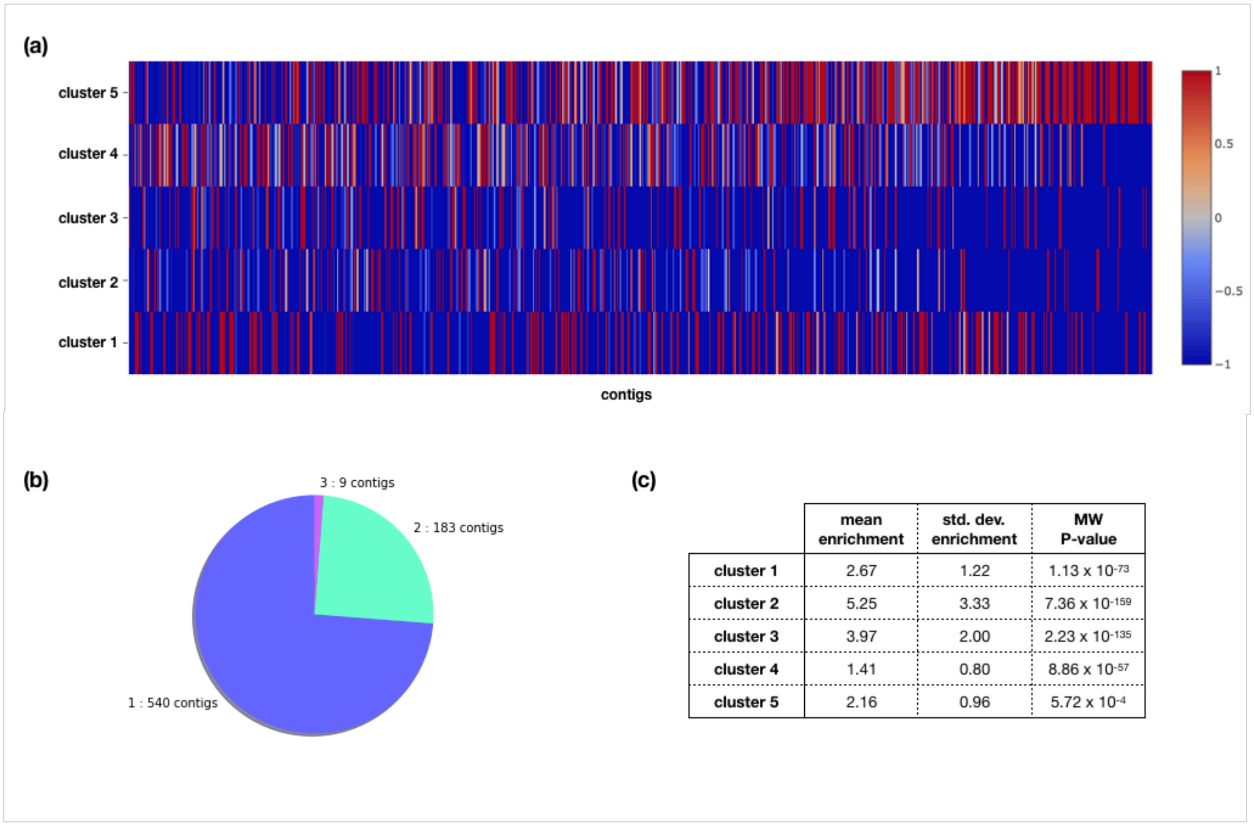
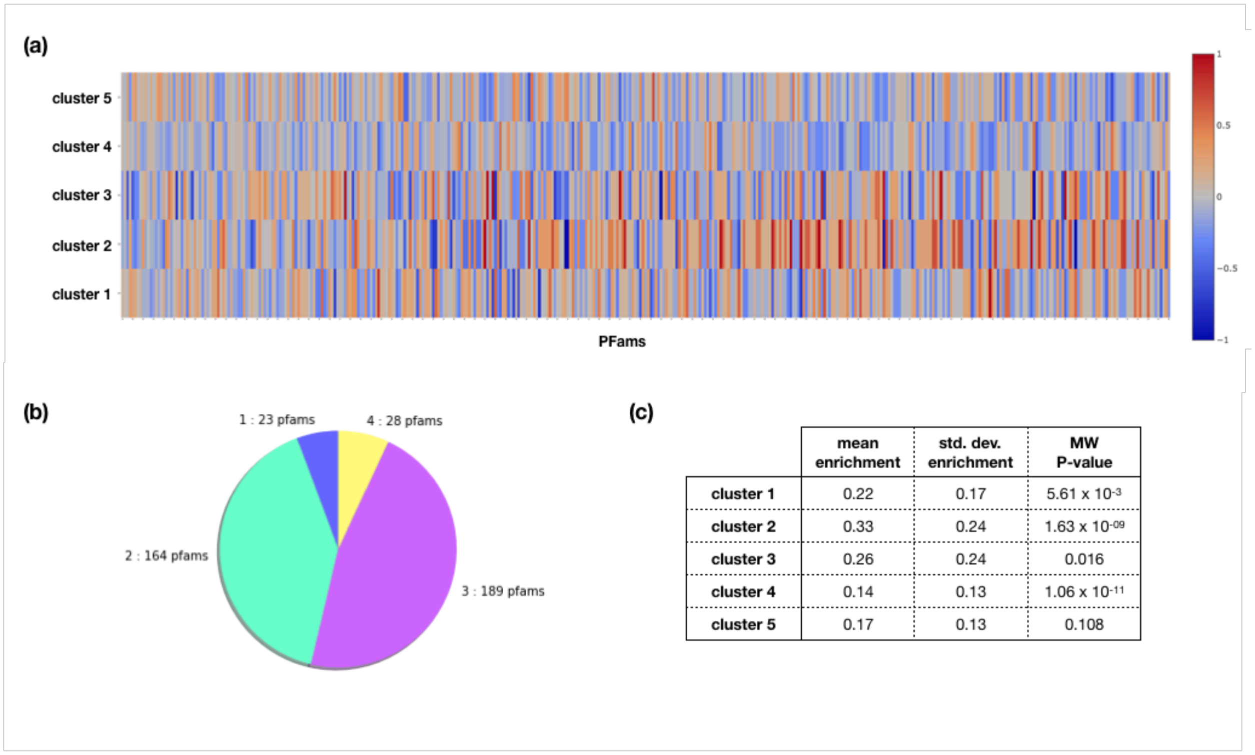


Fig. 3.4 – Relative proportion of clusters at each TARA site (vertical bars). TARA sites are sorted by longitude (x axis). Blue, cluster 1; green, cluster 2; yellow, cluster 3; purple, cluster 4; pink, cluster 5.



	mean enrichment	std. dev. enrichment	MW P-value
cluster 1	2.67	1.22	1.13×10^{-73}
cluster 2	5.25	3.33	7.36×10^{-159}
cluster 3	3.97	2.00	2.23×10^{-135}
cluster 4	1.41	0.80	8.86×10^{-57}
cluster 5	2.16	0.96	5.72×10^{-4}

Fig. 3.5 – Contig enrichment in clusters. (a) Heatmap of enrichment (red colors) or depletion (blue colors) of each SAR86 co-assembly contig (columns) across each cluster (rows). (b) Pie chart of the number of clusters in which SAR86 contigs are enriched. All contigs are enriched in three or fewer clusters, and most are enriched in only one. (c) Mean positive enrichment value, standard deviation of positive enrichment values, and the P value for the Mann-Whitney U test for whether a cluster enrichment vector is significantly different from null expected values, for each cluster.



	mean enrichment	std. dev. enrichment	MW P-value
cluster 1	0.22	0.17	5.61×10^{-3}
cluster 2	0.33	0.24	1.63×10^{-09}
cluster 3	0.26	0.24	0.016
cluster 4	0.14	0.13	1.06×10^{-11}
cluster 5	0.17	0.13	0.108

Fig. 3.6 – Functional enrichment in clusters. (a) Heatmap of enrichment (red) or depletion (blue) of the 405 most abundant Pfam families (columns) across each cluster (rows). Pfams are ordered left to right by the number of genes annotated to it, from the most abundant Pfams to the Pfams with as few as 20 genes annotated to it. (b) Pie chart of the number of clusters in which Pfams are enriched. (c) Mean positive enrichment value, standard deviation of positive enrichment values, and the P value associated with the Mann-Whitney U test for whether a cluster enrichment vector is significantly different from null expected values, for each cluster.

REFERENCES

Aguilar D, Aviles FX, Querol E, Sternberg MJE. Analysis of phenetic trees based on metabolic capabilites across the three domains of life. *J Mol Biol* 2004; **340**: 491–512.

Azam F. Microbial control of oceanic carbon flux: The plot thickens. *Science (80-)* 1998; **280**: 694–696.

Britschgi TB, Giovannoni SJ. Phylogenetic analysis of a natural marine bacterioplankton population by rRNA gene cloning and sequencing. *Appl Environ Microbiol* 1991; **57**: 1707–1713.

Carlson CA, Morris R, Parsons R, Treusch AH, Giovannoni SJ, Vergin K. Seasonal dynamics of SAR11 populations in the euphotic and mesopelagic zones of the northwestern Sargasso Sea. *ISME J* 2009; **3**: 283–295.

Cohan FM. Towards a conceptual and operational union of bacterial systematics, ecology, and evolution. *Philos Trans R Soc B Biol Sci* 2006; **361**: 1985–1996.

Delong EF, Preston CM, Mincer T, Rich V, Hallam SJ, Frigaard N, et al. Community Genomics Among Stratified Microbial Assemblages in the Ocean's Interior. *Science (80-)* 2006; **311**: 496–503.

Dupont CL, Rusch DB, Yooseph S, Lombardo MJ, Alexander Richter R, Valas R, et al. Genomic insights to SAR86, an abundant and uncultivated marine bacterial lineage. *ISME J* 2012; **6**: 1186–1199.

Dupont CL, Larsson J, Yooseph S, Ininbergs K, Goll J, Asplund-Samuelsson J, et al. Functional tradeoffs underpin salinity-driven divergence in microbial community composition. *PLoS One* 2014; **9**: e89549.

Falkowski PG, Fenchel T, Delong EF. The microbial engines that drive Earth's biogeochemical cycles. *Science* 2008; **320**: 1034–9.

Finn RD, Coggill P, Eberhardt RY, Eddy SR, Mistry J, Mitchell AL, et al. The Pfam protein families database: Towards a more sustainable future. *Nucleic Acids Res* 2016; **44**: D279–D285.

Galand PE, Pereira O, Hochart C, Auguet JC, Debroas D. A strong link between marine microbial community composition and function challenges the idea of functional redundancy. *ISME J* 2018; 1.

Guidi L, Chaffron S, Bittner L, Eveillard D, Larhlimi A, Roux S, et al. Plankton networks driving carbon export in the oligotrophic ocean. *Nature* 2015; **532**: in review.

Hoarfrost A. SAR86. *Github repository* 2018; https://github.com/ahoarfrost/SAR86/

Hunt DE, David LA, Gevers D, Preheim SP, Alm EJ, Polz MF. Resource partitioning and sympatric differentiation among closely related bacterioplanktn. *Science (80-)* 2008; **320**: 1081–1085.

Jiang X, Langille MGI, Neches RY, Elliot M, Levin S a., Eisen J a, et al. Functional Biogeography of Ocean Microbes Revealed through Non-Negative Matrix Factorization. *PLoS One* 2012; **7**: 1–9.

Kent AG, Dupont CL, Yooseph S, Martiny AC. Global biogeography of Prochlorococcus genome diversity in the surface ocean. *ISME J* 2016; **10**: 1856–1865.

Ladau J, Sharpton TJ, Finucane MM, Jospin G, Kembel SW, O'Dwyer J, et al. Global marine bacterial diversity peaks at high latitudes in winter. *ISME J* 2013; **7**: 1669–77.

Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Res* 2014; **42**: 490–495.

Louca S, Parfrey LW, Doebeli M. Decoupling function and taxonomy in the global ocean microbiome. *Science (80-)* 2016a; **353**: 1272–1277.

Louca S, Jacques SMS, Pires APF, Leal JS, Srivastava DS, Parfrey LW, et al. High taxonomic variability despite stable functional structure across microbial communities. *Nat Ecol Evol* 2016b; **1**: 0015.

Louca S, Polz MF, Mazel F, Albright MBN, Huber JA, O'Connor MI, et al. Function and functional redundancy in microbial systems. *Nat Ecol Evol* 2018.

Martiny JBH, Bohannan BJM, Brown JH, Colwell RK, Fuhrman JA, Green JL, et al. Microbial biogeography: putting microorganisms on the map. *Nat Rev Microbiol* 2006; **4**: 102–12.

Martiny JBH, Jones SE, Lennon JT, Martiny AC. Microbiomes in light of traits: A phylogenetic perspective. *Science (80-)* 2015; **350**.

Miller JR, Delcher AL, Koren S, Venter E, Walenz BP, Brownley A, et al. Aggressive assembly of pyrosequencing reads with mates. *Bioinformatics* 2008; **24**: 2818–2824.

Morales SE, Holben WE. Linking bacterial identities and ecosystem processes: Can 'omic' analyses be more than the sum of their parts? *FEMS Microbiol Ecol* 2011; **75**: 2–16.

Nayfach S, Rodriguez-Mueller B, Garud N, Pollard KS. An integrated metagenomics pipeline for strain profiling reveals novel patterns of bacterial transmission and biogeography. *Genome Res* 2016; **26**: 1612–1625.

Noguchi H, Park J, Takagi T. MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. *Nucleic Acids Res* 2006; **34**: 5623–5630.

Pommier T, Canbäck B, Riemann L, Boström KH, Simu K, Lundberg P, et al. Global patterns of diversity and community structure in marine bacterioplankton. *Mol Ecol* 2007; **16**: 867–80.

Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. *Mol Syst Biol* 2014; **7**.

Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S, et al. The Sorcerer II Global Ocean Sampling expedition: Northwest Atlantic through eastern tropical Pacific. *PLoS Biol* 2007; **5**: 0398–0431.

Rusch DB, Lombardo M-J, Yee-Greenbaum J, Novotny M, Brinkac LM, Lasken RS, et al. Draft genome sequence of a single cell of SAR86 clade subgroup IIIa. *Genome Announc* 2013; **1**: e00030-12.

Shi Y, Tyson GW, Eppley JM, Delong EF. Integrated metatranscriptomic and metagenomic analyses of stratified microbial assemblages in the open ocean. *ISME J* 2011; **5**: 999–1013.

Strickland MS, Lauber C, Fierer N, Bradford MA. Testing the functional significance of microbial community composition. *Ecology* 2009; **90**: 441–451.

Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. *Science (80-)* 2015; **348**: 1–10.

Suzuki MT, Beja O, Taylor LT, Delong EF. Phylogenetic analysis of ribosomal RNA operons from uncultivated coastal marine bacterioplankton. *Environ Microbiol* 2001; **3**: 323–331.

Swan BK, Tupper B, Sczyrba A, Lauro FM, Martinez-Garcia M, González JM, et al. Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proc Natl Acad Sci U S A* 2013; **110**: 11463–8.

Treseder KK, Balser TC, Bradford MA, Brodie EL, Dubinsky EA, Eviner VT, et al. Integrating microbial ecology into ecosystem models: Challenges and priorities. *Biogeochemistry* 2012; **109**: 7–18.

Treusch AH, Vergin KL, Finlay LA, Donatz MG, Burton RM, Carlson CA, et al. Seasonality and vertical structure of microbial communities in an ocean gyre. *ISME J* 2009; **3**: 1148–1163.

Tyberghein L, Verbruggen H, Pauly K, Troupin C, Mineur F, De Clerck O. Bio-ORACLE: A global environmental dataset for marine species distribution modelling. *Glob Ecol Biogeogr* 2012; **21**: 272–281.

Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, et al. Microbial species delineation using whole genome sequences. *Nucleic Acids Res* 2015; **43**: 6761–6771.

Wattam AR, Abraham D, Dalay O, Disz TL, Driscoll T, Gabbard JL, et al. PATRIC, the bacterial bioinformatics database and analysis resource. *Nucleic Acids Res* 2014; **42**: 581–591.

Wieder WR, Allison SD, Davidson EA, Georgiou K, Hararuk O, He Y, et al. Explicitly representing soil microbial processes in Earth system models. *Global Biogeochem Cycles* 2015; **29**: 1782–1800.

Widder S, Allen RJ, Pfeiffer T, Curtis TP, Wiuf C, Sloan WT, et al. Challenges in microbial ecology: building predictive understanding of community function and dynamics. *ISME J* 2016.

Zinger L, Amaral-Zettler L a, Fuhrman JA, Horner-Devine MC, Huse S, Welch DBM, et al. Global patterns of bacterial beta-diversity in seafloor and seawater ecosystems. *PLoS One* 2011; **6**: e24570.

CHAPTER 4: SEQUENCING DATA DISCOVERY WITH METASEEK

Adrienne Hoarfrost[1,*], Nick Brown[2], C. Titus Brown[3], and Carol Arnosti[1]

[1]Department of Marine Sciences, University of North Carolina at Chapel Hill, Chapel Hill, NC,

[2]Independent Researcher, Durham, NC,

[3]Department of Veterinary Medicine, University of California Davis.

*To whom correspondence should be addressed.

In submission to *Bioinformatics*

1 Introduction

Sequencing data generation is rapidly increasing, as of 2018 reaching more than 3 million sequencing datasets in the Sequence Read Archive (SRA), the primary repository for next-generation sequencing data in the International Nucleotide Sequence Database Collaboration (NCBI 2017). As research communities produce data at increasingly rapid rates, there is growing interest in leveraging these data resources for new insights into biological systems using comparative meta-analyses of large-scale integrated datasets. Data curation is the first step in this process, and generally requires identifying datasets in data repositories that match certain criteria that may be described within the datasets' metadata.

However, easy-to-use, flexible, and comprehensive tools for searching and filtering existing data repositories according to their metadata parameters are lacking. The e-utilities tool provided by the National Center for Biotechnology Information (NCBI), for example, is restricted to a free text search or exact string matching on a limited set of fields (Sayers 2017). A tool such as SRAdb (Zhu *et al.*, 2013), in contrast, expands the searchability of metadata fields, but is specific to the R programming language and requires a local build of the SRAdb database. Neither of these tools, meanwhile, address the widespread errors in sequencing metadata, which is collected mainly via user-provided free text entries that result in frequent misspellings, missing fields, and nonobservance of existing metadata standards, the Minimum Information about any Sequence (MIxS) specification (Yilmaz et al. 2011).

MetaSeek provides a sequencing data discovery tool that facilitates easy and rapid curation of integrated sequencing datasets. The MetaSeek interface is intuitive, user-friendly, and flexible, allowing users to search on any metadata field in any of 10 ways. For programmatic access, MetaSeek exposes a simple API that is programming language agnostic.

2 Infrastructure & Implementation

MetaSeek automatically scrapes metadata from the SRA on a weekly basis. In the SRA and in MetaSeek, each '#RX' accession ID (SRX, ERX, or DRX depending on whether it originates from the USA NCBI, European EBI, or Japanese DDBJ databases respectively) is a unique metadata entry. Metadata for each dataset are gathered from across the SRA, BioSample, and PubMed databases and unified for each MetaSeek dataset entry.

As metadata are scraped from the SRA, they are cleaned and parsed to be compliant with MIxS metadata standards. Redundant fields are unified into a single field name, while fields with categorical inputs are parsed where possible to these values, rather than free text entries. Numerical fields that are gathered as free text, such as latitude and longitude, are parsed into numeric values as well. Finally, some fields with commonly missing metadata can be inferred from the other metadata context: investigation_type, an essential MIxS standard field, is often not provided by the user but can be predicted by logistic regression with 94.1% accuracy from the library_source, library_strategy, library_screening_strategy, and study_type fields.

The cleaned metadata are stored in the MetaSeek database, which is wrapped with an API implemented in Python's Flask library. The API interfaces communication between the database server and the MetaSeek web front-end, which is implemented in React, a popular JavaScript library for interactive web applications.

The MetaSeek database is hosted by Amazon Redshift, while the API and front-end is hosted on an Amazon EC2 server. Redshift provides a columnar data storage schema that allows for rapid response times to analytical workloads such as the summary histogram counts of metadata fields seen in the MetaSeek "Explore" page.

3 Interfacing With MetaSeek

MetaSeek search, filter, and download functionality can be accessed via both the interactive online interface and a programmatic API. While the web interface emphasizes ease of use, the API emphasizes flexibility and comprehensiveness. Together, the online interface and API meet the needs of both casual and in-depth users.

3.1 The Online Interface

The main search and filter functionality is provided on the "Explore" page (www.metaseek.cloud/explore). A filter panel provides intuitive filter options for the most useful

MetaSeek database fields. As users enter filter parameters, summary information of the datasets matching these filter parameters, such as counts of categorical variables, histograms of numeric fields, and a geographic map of dataset origins, are shown in real-time in an interactive visualization dashboard.

Users can save their configured filter parameters as “discoveries”, where they are accessible at a later date, shareable with other users, or able to be referenced in a publication. From the “Discovery Details” page, .csv files of either matching dataset IDs or all available metadata for each matching dataset can be downloaded directly. All user discoveries are made public and can be browsed on the “Browse” page, where previously saved discoveries can be used as a launching off point for other users.

3.2 The MetaSeek API

The MetaSeek API provides a programmatic interface for querying the MetaSeek database. It is programming language agnostic and can be accessed via any HTTP POST request. The core API calls, SearchDatasetIds and SearchDatasetMetadata, take a set of filter parameters as input and return either a list of matching dataset IDs (SearchDatasetIds) or the full metadata (SearchDatasetMetadata) for every matching dataset. Filter parameters are flexible, such that any field in the MetaSeek database can be filtered by any value provided by the user, in any of 10 ways called “rule types”. Rule types are indicated by the user by an integer corresponding to the desired rule type, which consist of: “greater than”, “less than”, “greater than or equal to”, “less than or equal to”, “is equal to”, “is *not* equal to”, “is equal to any of a list of items”, “is *not* equal to any of a list of items”, “contains the partial text”, and “is not null”.

4 Conclusions

MetaSeek fills a growing need in the bioinformatics community for faster, easier, and more accurate data discovery and integration. Future development will focus on curation of metadata from additional sequencing data repositories, direct integration with bioinformatics tools, and metadata inference from unstructured text data. Feature requests and input from the community are welcome and can be submitted via the website or directly to metaseek.cloud@gmail.com. Future updates and feature additions will be announced on the MetaSeek website.

Funding

This work has been supported by Gordon and Betty Moore Foundation GBMF4551 to CTB, NSF OCE-1736772 to CA, and a Deep Carbon Observatory Deep Life Modeling & Visualization Graduate Fellowship to AH.

Conflict of Interest: none declared.

Figures

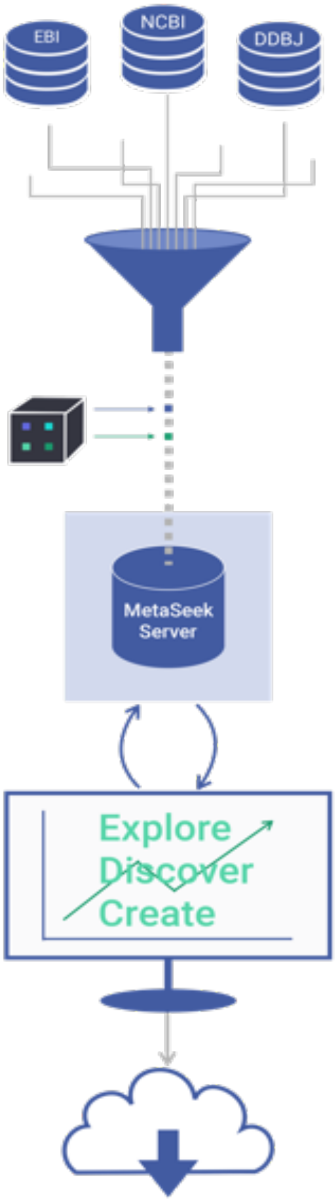


Fig 4.1 – The MetaSeek workflow.

REFERENCES

NCBI Resource Coordinators. (2013). Database resources of the National Center for Biotechnology Information. *Nucleic Acids Research*, *41*, D8–D20. http://doi.org/10.1093/nar/gks1189

Sayers E. (2017). *The E-utilities In-Depth: Parameters, Syntax and More*. 2009 May 29 [Updated 2017 Nov 1]. In: Entrez Programming Utilities Help [Internet]. Bethesda (MD): National Center for Biotechnology Information (US).

Yilmaz P, Kottmann R, Field D, Knight R, Cole JR, Amaral-Zettler L, et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol,* **29**, 415–420.

Zhu Y, Stephens RM, Meltzer PS, Davis SR. (2013). SRAdb: query and use public next-generation sequencing data from within R. *BMC Bioinformatics,* **14**(19).

DISCUSSION

Microbial communities interact with their environment to drive the functional and biogeochemical outcomes of ecosystems, while environmental variables in turn influence the distribution of microbial genetic diversity. Decades of scientific inquiry have increased our appreciation of the vast complexity and interconnectivity between microbial communities and their environment (Azam 1998, Falkowski et al. 2008, Lima-Mendez et al. 2015), but significant questions remain in our understanding of the impact of microbial processes on biogeochemical fluxes, environmental regulation of the distribution of microbial genetic diversity, and building effective models to predict these complex ecosystem-level outcomes. This dissertation demonstrates the relationship between physical processes and biogeochemical outcomes at the regional scale, and how these outcomes may change as environmental conditions shift in the future. These regional differences are also visible at the global scale, where the geographical distribution of SAR86 genetic diversity, and the global distribution of SAR86 ecotypes, can be accurately predicted from environmental variables. As our scale of inquiry expands from the community, to regional, and finally to global scales, so too does the complexity of microbe-environment interactions in the Earth system. Our ability to understand and predict these complex systems will depend on our ability to expand both the range of environments feasible for study at the community scale, as well as the computational limits of data-driven discovery at ever larger spatial, temporal, and computational scales.

The expansion of community-level feasibility is exemplified by the extraction treatment presented in Chapter 1. It is a popular adage that we know more about the surface of the moon than we do about the bottom of the ocean. The deep subsurface is even more underexplored, despite the fact that volumetrically it is the largest biome on Earth, inhabited by a microbial population of roughly 10^{29} cells (Kallmeyer et al. 2012) rivaling in number that of both seawater and soils (Whitman et al. 1998). Expanding our ability to measure microbial activities in sedimentary environments is key to increasing our understanding of the diversity of microbial systems, the range of environmental conditions they inhabit, and the diversity of microbial survival strategies at the edges of the limits of life. The method introduced in Chapter 1 improves fluorescent substrate recovery from sedimentary systems by 66%, expanding the range of environments for which microbial activity measurements are possible to some of the least well-studied environmental frontiers.

The results of Chapters 2 and 3 demonstrate the close association between microbial and environmental processes, as well as the potential for understanding and predicting microbial systems and their relationship with the environment with increasing complexity at increasing scales. The unique spectrum of carbon cycling rates and capacities of microbial communities in distinct water masses during eddy intrusions on the Mid Atlantic Bight has implications for biogeochemical cycling, with low bacterial productivity as well as a distinct spectrum of hydrolytic substrates characterizing warm core ringwater intrusions. As the oceans warm and similar eddy intrusions become more frequent (Andres 2016; Monim 2017; Gawarkiewicz et al. 2018), the influence of microbial communities within eddy intrusions on biogeochemical cycling along continental shelf regions is likely to increase as well. The low bacterial productivity and hydrolytic spectra unique to warm core eddies, despite warmer temperatures relative to surrounding water masses, demonstrates that predicting rates of biological carbon cycling in marine environments cannot be as straightforward as a simple temperature-growth relationship. Current Earth system models, however, either ignore biology completely or make simple assumptions about rates of activity due to temperature (Treseder et al. 2012; Wieder et al. 2015) that overlook the contribution of microbial community genetic capacities to their ultimate biogeochemical function, and the variable distribution of this genetic diversity in the world's oceans. However, it is also clear that there is a relationship between environmental variables and the distribution of genetic diversity, as demonstrated in Chapter 3, where SAR86 gene distributions could be predicted with high accuracy from environmental variables with machine learning models. These gene models enabled the identification of five ecotypes within SAR86 with distinct environmental distributions across the world's oceans. The differential taxonomic and functional diversity across ecotypes has implications for biogeochemical cycling; however, the link between this vast diversity and net biogeochemical outcomes is not straightforward. A significant challenge going forward will be in deriving biogeochemically and ecologically relevant features from complex microbial data, which can be used to more closely relate microbial systems to the environmental processes they regulate, and improve our predictions of ecosystem function under past, present, and future Earth conditions.

Deriving patterns and high-level features from microbial systems is an extremely data-intensive task. The MetaSeek data discovery tool provides a much needed resource to integrate datasets at the scale needed for ecosystem-scale investigations of microbe-environment interactions. As data size expands, our data analysis and modeling approaches must also adapt to best harness the potential of these data resources. Machine learning approaches, such as those used to identify SAR86 ecotypes from the global TARA dataset, are one such solution, but significant potential remains in the adaptation of these tools for bioinformatics applications (Min et al. 2017; Soueidan & Nikolski 2016).

The coming years in environmental microbiology will be defined by our ability to integrate knowledge across all levels of microbial organization, from the local environments they inhabit to the global-scale processes they regulate; to leverage data-intensive integrations to uncover patterns in the complexity inherent to these systems; and to use this information to formulate new and more comprehensive theories of microbial organization and their functional role within the Earth system. This dissertation addresses questions central to microbial-environmental interactions, linking microbial processes to environmental phenomena from community to global scales, and lays the groundwork for future investigation in data-driven discovery in environmental microbiology.

REFERENCES

Andres, M. 2016. On the recent destabilization of the Gulf Stream path downstream of Cape Hatteras. Geophys. Res. Lett. **43**. doi:10.1002/2016GL069966.

Azam F. (1998) Microbial control of oceanic carbon flux: The plot thickens. *Science;* **280**:694–696.

Falkowski PG, Fenchel T, Delong EF. (2008) The microbial engines that drive Earth's biogeochemical cycles. *Science*; **320**:1034–9.

Gawarkiewicz G, Todd RE, Zhang W, Partida J, Gangopadhyay A, Monim M-U-H, et al. (2018) The changing nature of shelfbreak exchange revealed by the OOI Pioneer Array. *Oceanography*; **31**:60–70.

Kallmeyer J, Pockalny R, Adhikari RR, Smith DC, D'Hondt S. (2012) Global distribution of microbial abundance and biomass in subseafloor sediment. *Proc Natl Acad Sci*; **109**:16213–6.

Lima-Mendez, G. et al. (2015) Determinants of community structure in the global plankton interactome. *Science*; 348.

Min S, Lee B, Yoon S. (2017) Deep learning in bioinformatics. *Brief Bioinform*; **18**:851–869.

Monim, M. 2017. Seasonal and Inter-annual Variability of Gulf Stream Warm Core Rings from 2000 to 2016. University of Massachusetts-Dartmouth.

Soueidan H, Nikolski M. (2016) Machine learning for metagenomics: methods and tools. *arXiv*; 1–19.

Treseder KK, Balser TC, Bradford MA, Brodie EL, Dubinsky EA, Eviner VT, et al. (2012) Integrating microbial ecology into ecosystem models: Challenges and priorities. *Biogeochemistry*; **109**: 7–18.

Whitman WB, Coleman DC, Wiebe WJ (1998) Prokaryotes: The unseen majority. *Proc Natl Acad Sci;* **95**:6578–6583.

Wieder WR, Allison SD, Davidson EA, Georgiou K, Hararuk O, He Y, et al. (2015) Explicitly representing soil microbial processes in Earth system models. *Global Biogeochem Cycles*; **29**: 1782–1800.

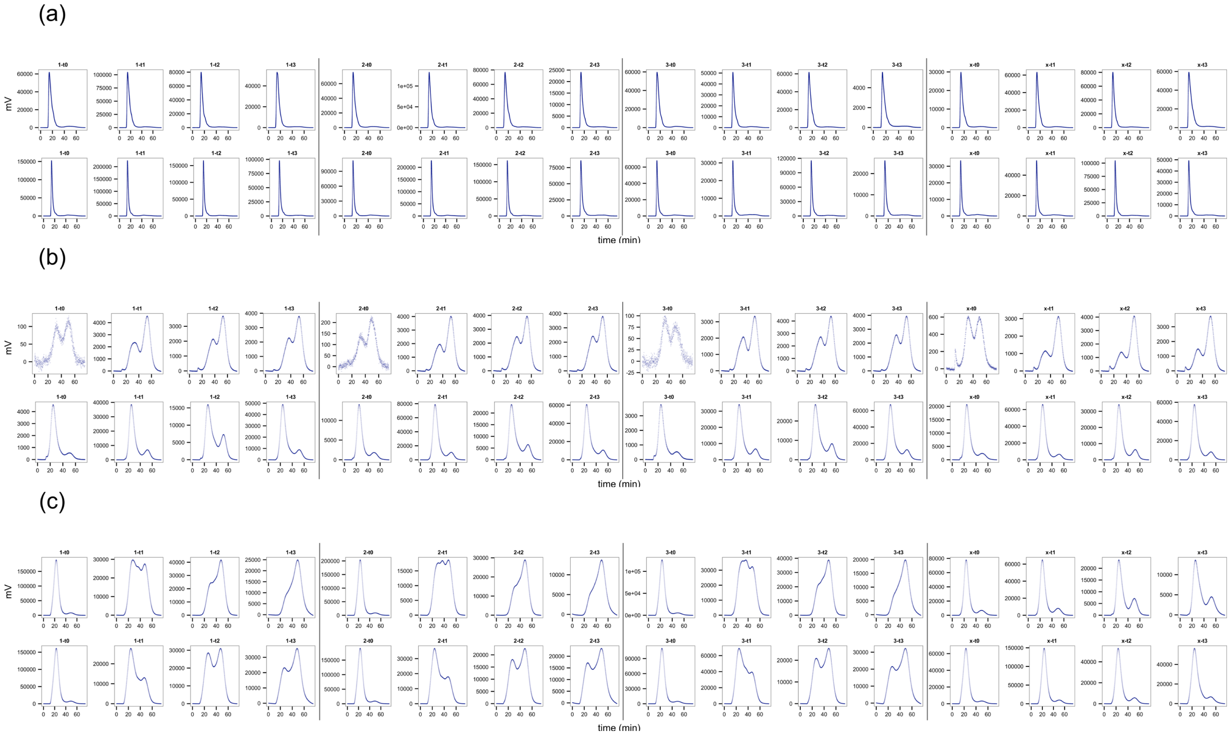
APPENDIX A: CHAPTER 1 SUPPLEMENTARY INFORMATION

Supplementary Tables

Sediment Source	Sediment Description
Marmara Sea	Marmara Sea sediments were dominated by clay sediments (>50%) at the surface, and were rich in micro-carbonates containing abundant (~10%) coccoliths. Marls of primarily carbonate mudstone characterized the deeper sediments from 520 and 570cm, which also contained abundant small sulfide particles (Zabel et al., 2011).
Eastern Mediterranean	Eastern Mediterranean sediments contained five sapropel layers that were cross-referenced with those described by Calvert and Fortugne (2001). Those used in these experiments included S4 (from 385cm), S5 (455cm), and S7 (582-590cm), which were all dark brown/black, organic-rich clays. The remaining non-sapropelic layers were gray to light-gray coccolith ooze and clay (Zabel et al., 2011).
Guaymas Basin	Cores P1 and P3 consisted of diatom-rich, non-sulfidic hemipelagic marine sediments, and were both collected on the outermost northwestern ridge flank of Guaymas Basin (27°38.27' N, 111°53.89' W; 27°37.68' N, 111°52.57' W) at water depths of 1604 m and 1611m, respectively. Core P5 (27°38.76' N, 111°38.91' W) was sampled at the foot of the Sonora Margin, and contained highly compacted and sulfidic sediments. Core P8 contained suboxic sediments and was sampled on the upper Sonora Margin (27°40.34' N, 111°24.12' W) at 995 m water depth. Core P10, dominated by hemipelagic, diatom-rich sediments was sampled in the central region of the northwestern ridge flanks (27°30.52' N, 111°42.17' W) at 1731 m water depth. Core P13 was sampled on the southeastern edge of the ridge flank (27°12.45' N, 111°13.77' W), on the opposite side of the axis from P1 and P3. Sediments sampled from this core showed strong terrestrial impacts from a nearby river delta, with coarse and sandy sediments and sand layers.

SI Table 1.1 – Description of sediments collected from the three sites used in this study – Marmara Sea, Eastern Mediterranean, and Guaymas Basin.

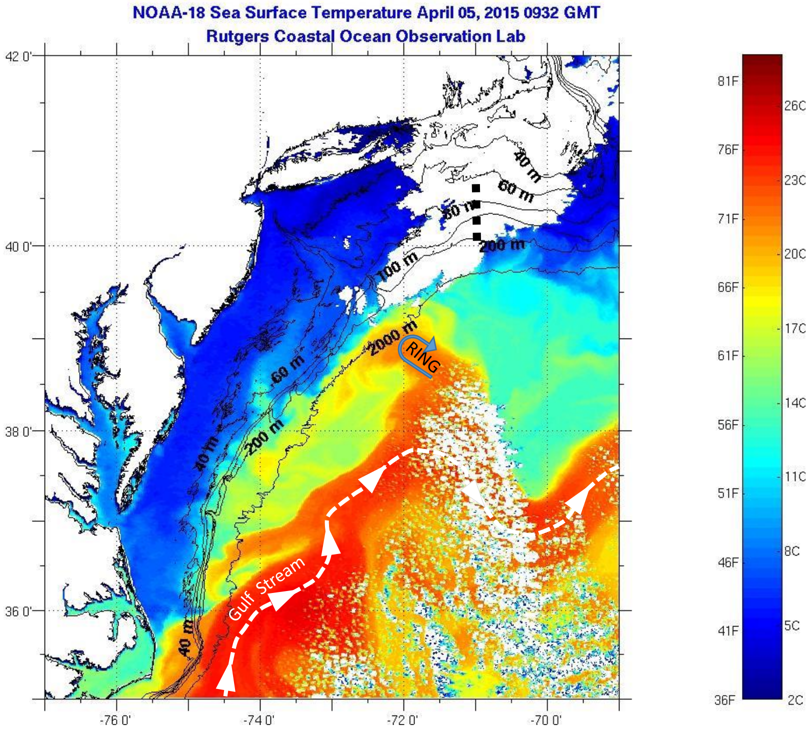
Supplementary Figures



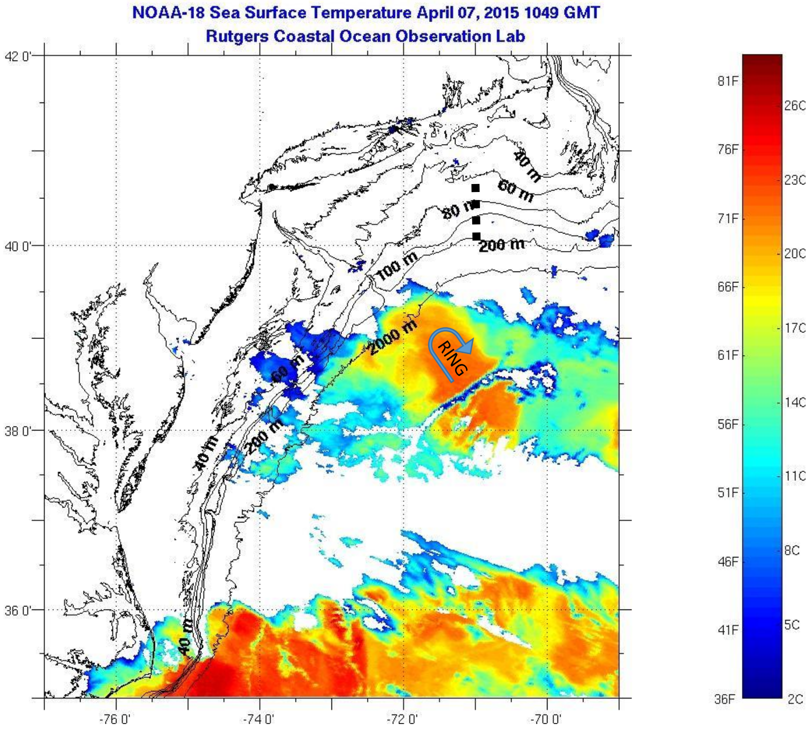
SI Fig 1.1 - Representative chromatograms comparing treatment (bottom row) to no treatment controls (top row) for (a) Mediterranean sapropel S4, 385cm, chondroitin incubations (b) Guaymas core P13, 55cm, laminarin, and (c) Guaymas core P1, 55cm, laminarin. Note differences in scales on y axes. Replicate 1, far left vertical panels, is duplicated from Fig 2. Improved chromatogram quality is seen in treatment incubations, with narrower peak widths, higher total integrated fluorescence, and a higher proportion of high- to low-molecular-weight substrate.

APPENDIX B: CHAPTER 2 SUPPLEMENTARY INFORMATION

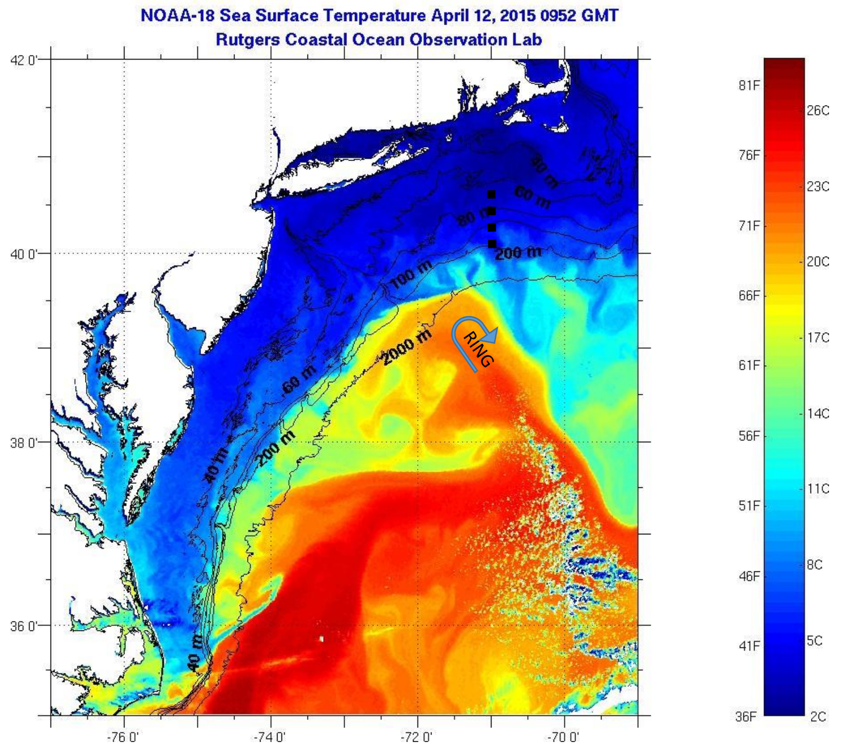
Supplementary Figures

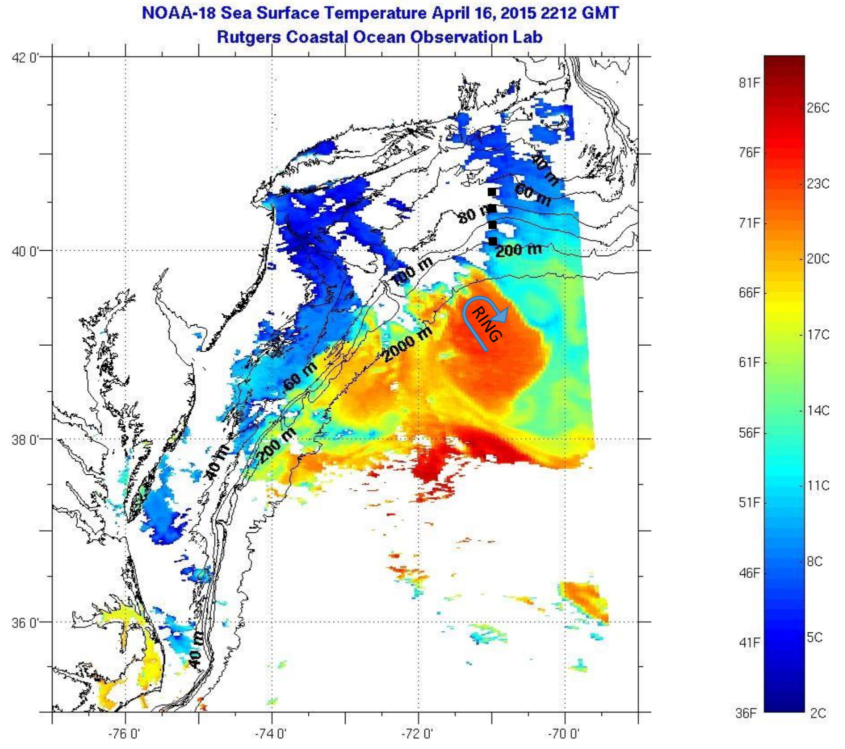


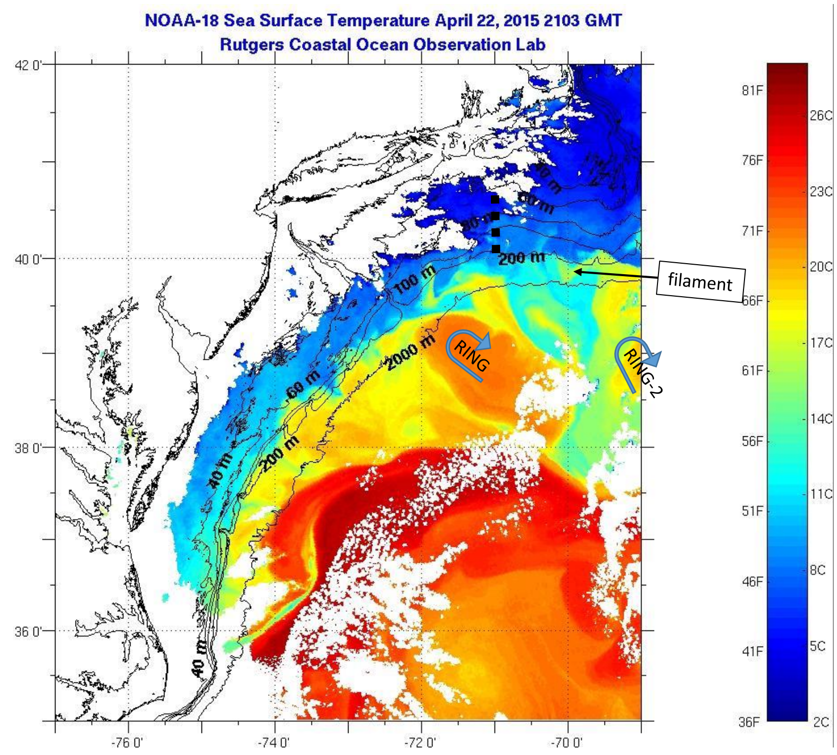
NOAA-18 Sea Surface Temperature April 05, 2015 0932 GMT
Rutgers Coastal Ocean Observation Lab

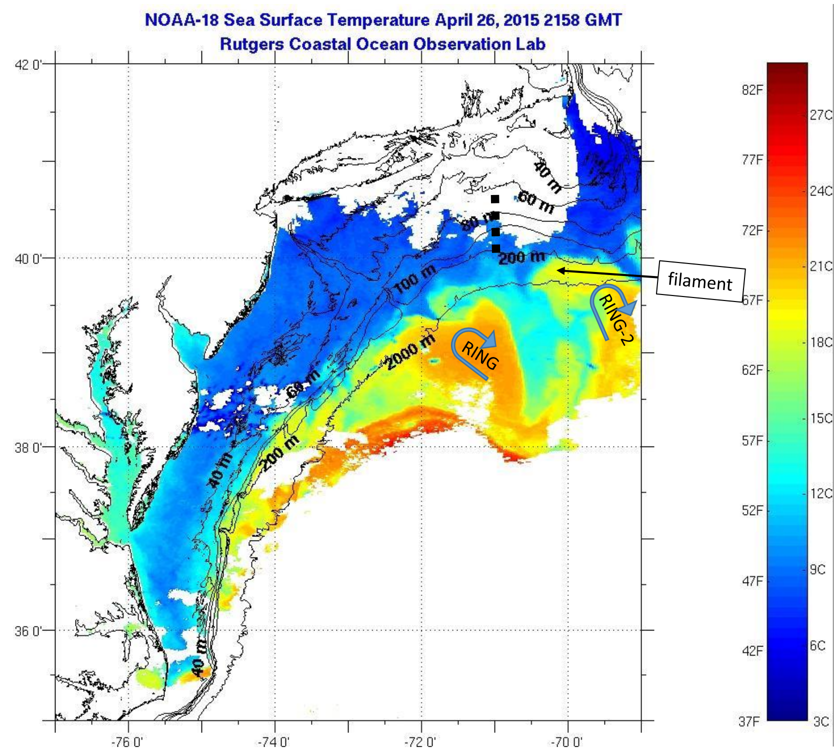


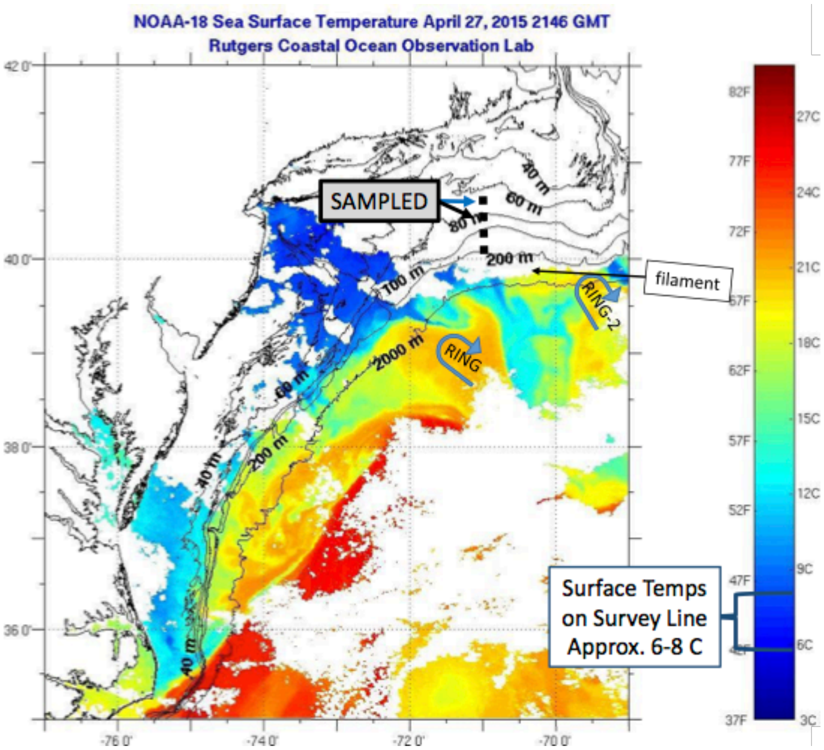
NOAA-18 Sea Surface Temperature April 07, 2015 1049 GMT
Rutgers Coastal Ocean Observation Lab

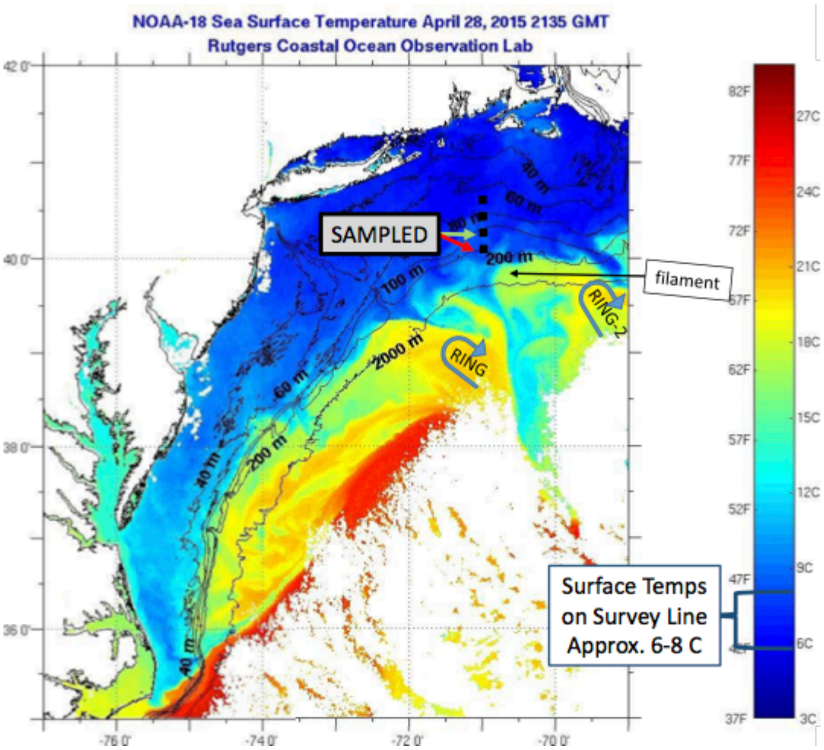

NOAA-18 Sea Surface Temperature April 12, 2015 0952 GMT
Rutgers Coastal Ocean Observation Lab
42 0'
40 0'
38 0'
36 0'
-76 0'
-74 0'
-72 0'
-70 0'
40 m
60 m
80 m
200 m
100 m
2000 m
RING
81F
76F
71F
66F
61F
56F
51F
46F
41F
36F
26C
23C
20C
17C
14C
11C
8C
5C
2C

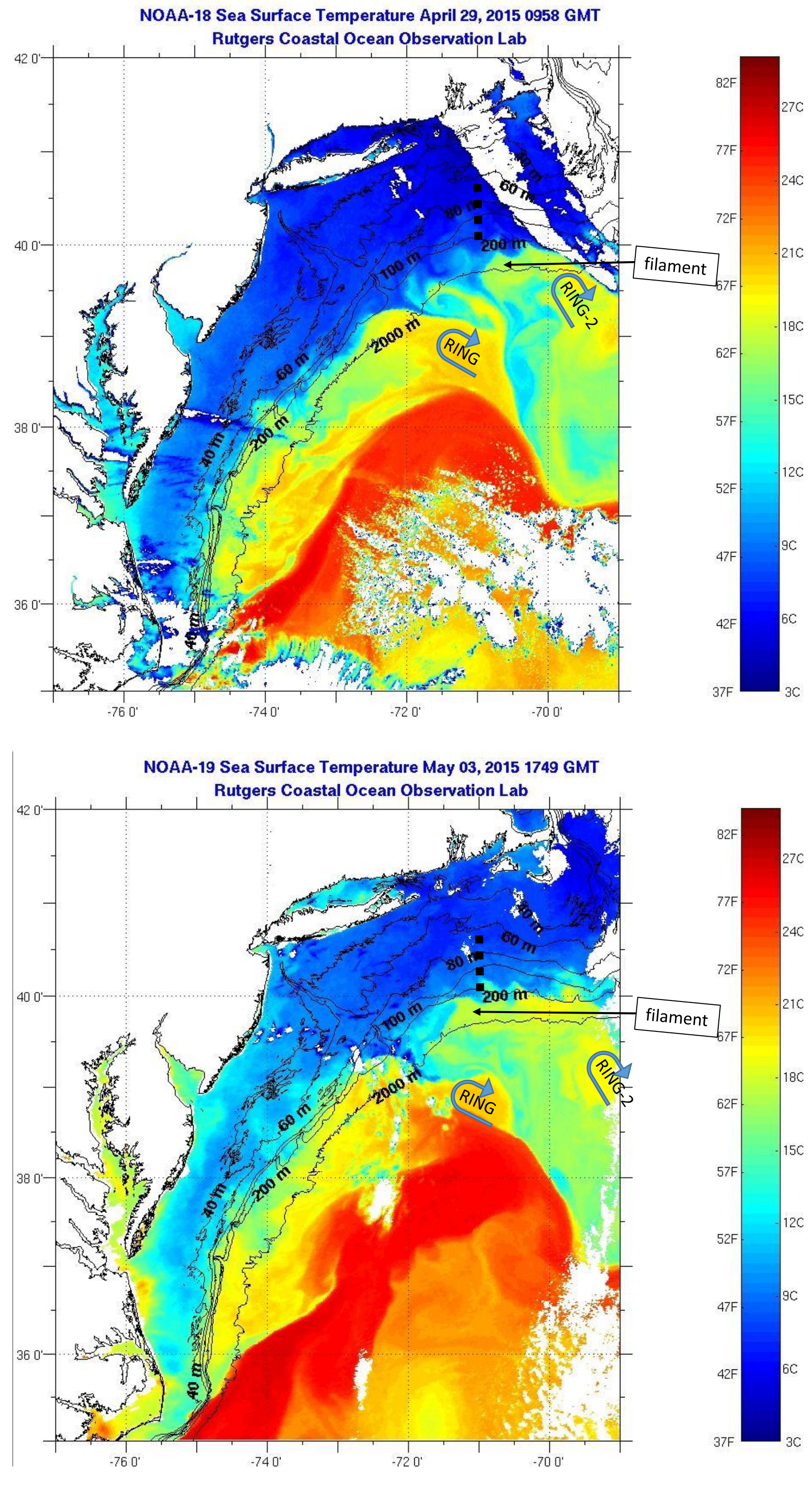

NOAA-18 Sea Surface Temperature April 16, 2015 2212 GMT
Rutgers Coastal Ocean Observation Lab
42 0'
40 0'
38 0'
36 0'
-76 0'
-74 0'
-72 0'
-70 0'
40 m
60 m
80 m
200 m
100 m
2000 m
RING
81F
76F
71F
66F
61F
56F
51F
46F
41F
36F
26C
23C
20C
17C
14C
11C
8C
5C
2C


NOAA-18 Sea Surface Temperature April 22, 2015 2103 GMT
Rutgers Coastal Ocean Observation Lab
42 0'
40 0'
38 0'
36 0'
-76 0'
-74 0'
-72 0'
-70 0'
40 m
60 m
100 m
200 m
2000 m
RING
RING-2
filament
81F
76F
71F
66F
61F
56F
51F
46F
41F
36F
26C
23C
20C
17C
14C
11C
8C
5C
2C

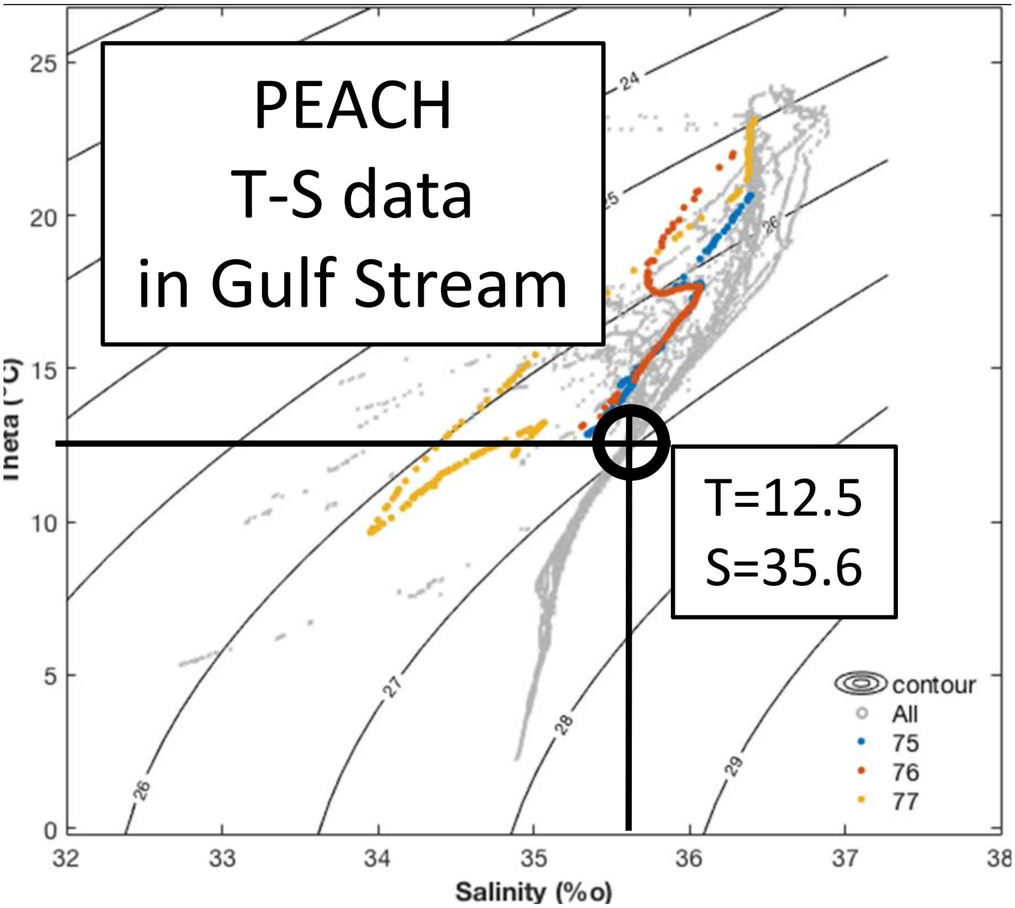

NOAA-18 Sea Surface Temperature April 26, 2015 2158 GMT
Rutgers Coastal Ocean Observation Lab
42 0'
40 0'
38 0'
36 0'
-76 0'
-74 0'
-72 0'
-70 0'
40 m
60 m
100 m
200 m
2000 m
RING
RING-2
filament
82F
77F
72F
67F
62F
57F
52F
47F
42F
37F
27C
24C
21C
18C
15C
12C
9C
6C
3C


NOAA-18 Sea Surface Temperature April 27, 2015 2146 GMT
Rutgers Coastal Ocean Observation Lab
SAMPLED
filament
RING
RING-2
Surface Temps on Survey Line Approx. 6-8 C
40 m
60 m
80 m
100 m
200 m
2000 m
42 0'
40 0'
38 0'
36 0'
-76 0'
-74 0'
-72 0'
-70 0'
82F
77F
72F
67F
62F
57F
52F
47F
37F
27C
24C
21C
18C
15C
12C
9C
6C
3C


NOAA-18 Sea Surface Temperature April 28, 2015 2135 GMT
Rutgers Coastal Ocean Observation Lab
SAMPLED
filament
RING
RING-2
Surface Temps on Survey Line Approx. 6-8 C
30 m
40 m
60 m
80 m
100 m
200 m
2000 m
42 0'
40 0'
38 0'
36 0'
-76 0'
-74 0'
-72 0'
-70 0'
82F
77F
72F
67F
62F
57F
52F
47F
37F
27C
24C
21C
18C
15C
12C
9C
6C
3C



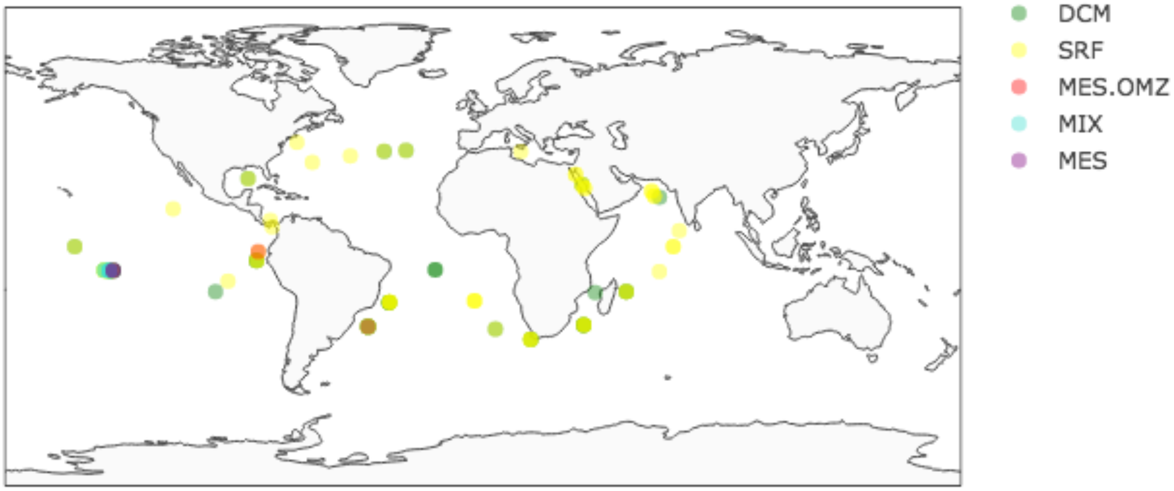
SI Fig. 2.1 – Sea surface temperature from satellite imagery of the Rutgers Coastal Ocean Observation Lab for April 5 to May 3, 2015. Station sampling locations are denoted by black squares along ~70°W longitude in each image. The path of the Gulf Stream jet is indicated by white arrows on April 5. The anticyclonic circulation around each of two Gulf Stream warm core rings is indicated with blue curved arrows whenever the rings are apparent. One warm core ring (RING) approaching the survey line from the west is observed during April 5 to 22. A second warm core ring (RING-2) is seen approaching the area from the southeast beginning on April 22. RING-2 moves close to the survey area by April 27-28, and a warm water filament is pushed toward the survey line ahead of it. The activity of these two rings likely brought warm/salty water of Gulf Stream origin to the upper continental slope at the southern end of the survey line. The cold/fresh water that can be seen in the April 27-28 images flowing southward from the survey line is likely the water measured in the upper ~20 meters at each of the 4 stations.



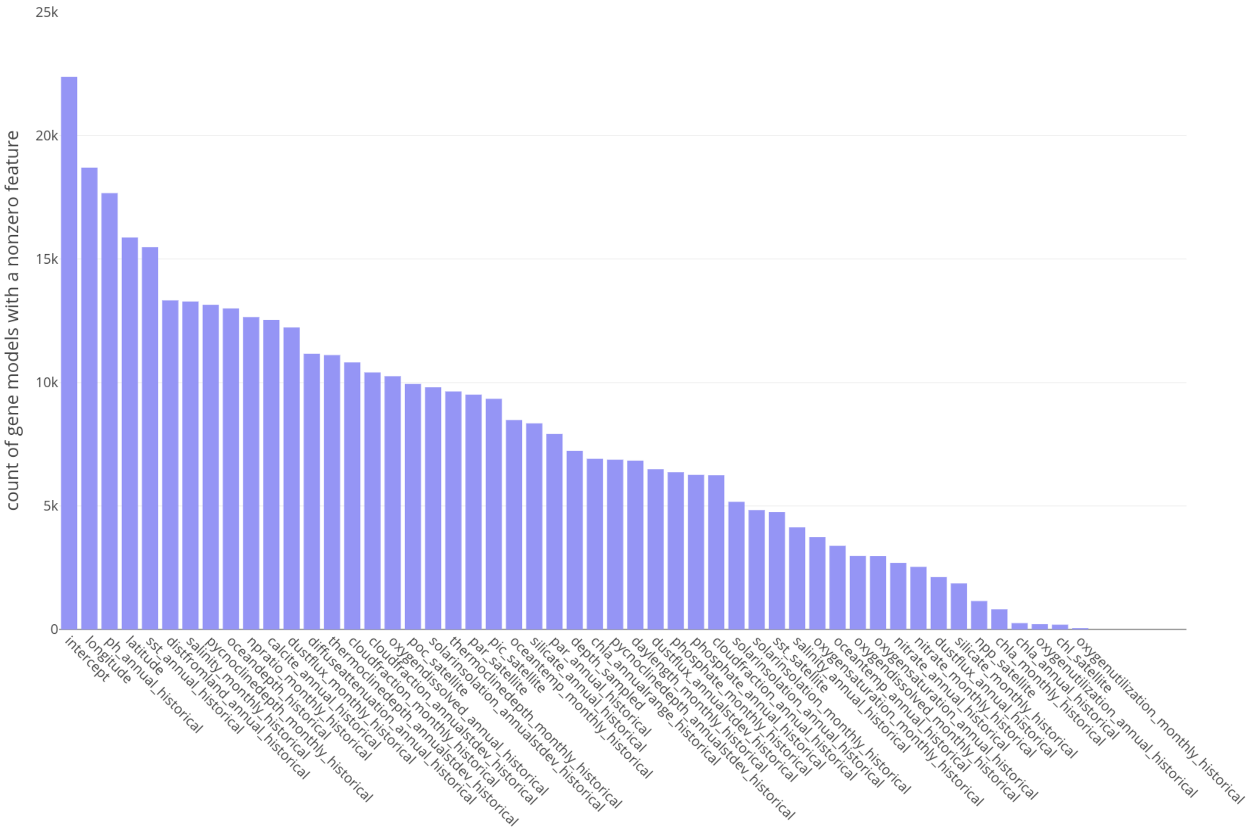
SI Fig. 2.2 – Recent observations of T-S characteristics of Gulf Stream waters, from the "Processes driving Exchange at Cape Hatteras" (PEACH) project. The T-S character (Temp ~12.5°C, Salinity ~35.6 PSU) of the water below 120 meters at station 4 is indicated on the plot, and is consistent with recent observations of water within the Gulf Stream jet at the southern perimeter of the Mid Atlantic Bight (see April 28 image of SI Fig 1), suggesting the deep water at Stn 4 is of Gulf Stream origin, likely brought to the survey area by warm core rings.

APPENDIX C: CHAPTER 3 SUPPLEMENTARY INFORMATION

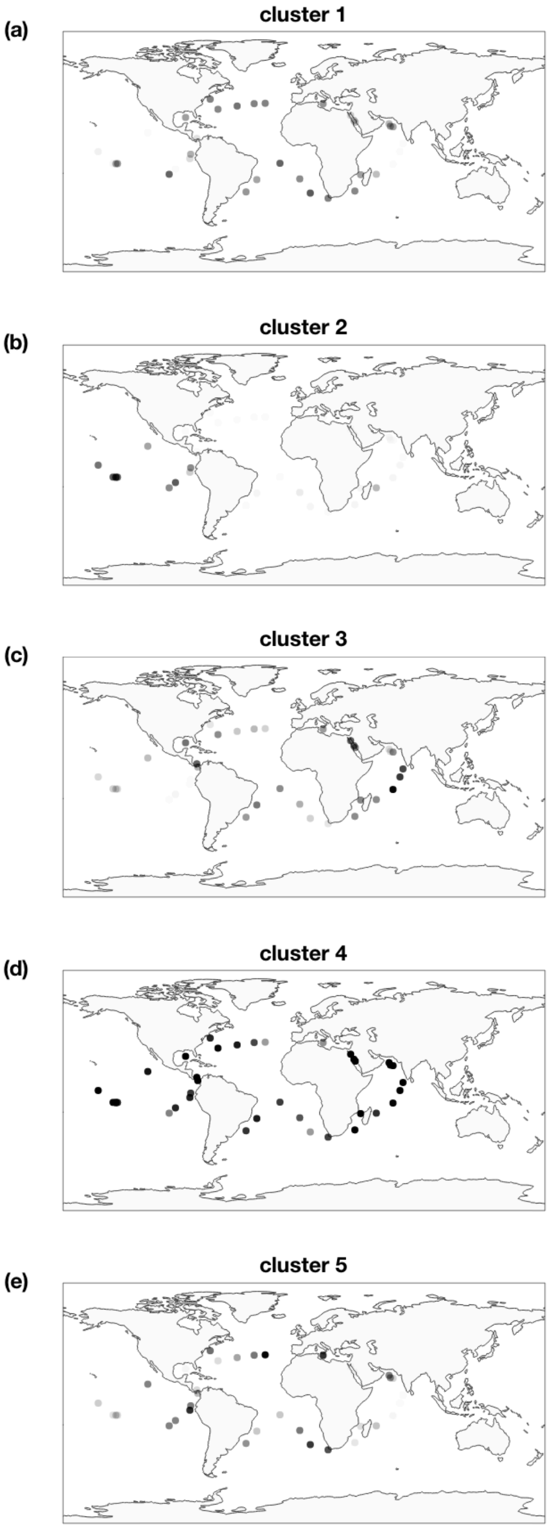
Supplementary Figures



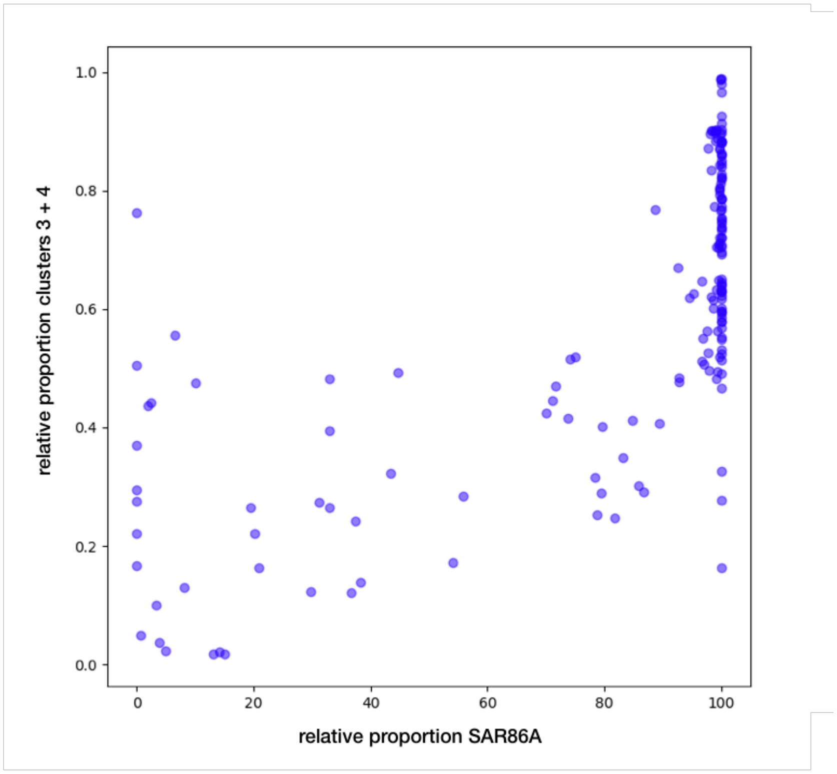
SI Fig. 3.1 – Map of TARA station sampling locations. Points are colored by depth region sampled: DCM: deep chlorophyll maximum; SRF: surface; MES.OMZ: mesopelagic in ocean minimum zone; MIX: mixotrophic; and MES: mesopelagic.



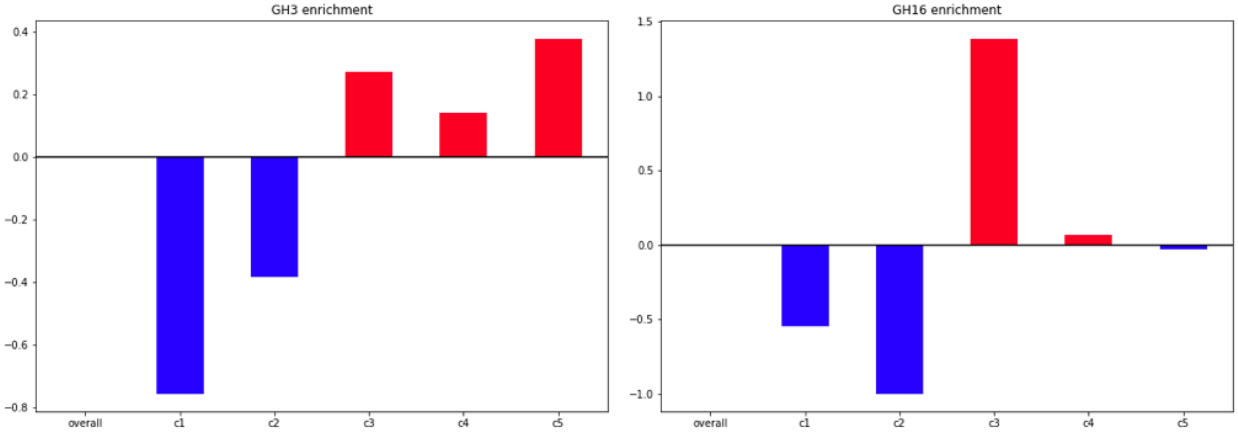
SI Fig. 3.2 – Number of gene models (y axis) for which a particular environmental feature (x axis) was selected as nonzero during logistic regression model training. Features are sorted in order of the number of gene models for which it was nonzero.



SI Fig. 3.3 – Map of the relative proportion of biogeographic clusters 1-5, top to bottom, at each TARA site. The opacity of each symbol indicates the percentage of genes from that TARA site that belong to that cluster.



SI Fig. 3.4 – Correlation between the relative abundance of SAR86A (vs SAR86E) and the relative proportion of clusters 3 + 4 (vs clusters 1 + 5) at each TARA site (blue dots). Pearson $R^2 = 0.70$, $P = 1.56 \times 10^{-26}$.



SI Fig. 3.5 – Functional enrichment of Pfams associated with glycosyl hydrolase family 3 (left) and glycosyl hydrolase family 16 (right) in each cluster. The expected value (‘overall’) is indicated by a horizontal line at 0, enriched values (red) appear above this line and depletion values (blue) below this line.

Supplementary Table Info, Captions, & Download

Because the supplementary tables described below are too large for print, the tables have been made available on the associated Github repository (Hoarfrost 2018). Tables can be downloaded directly from the url: https://github.com/ahoarfrost/SAR86/tree/master/SI_tables.

SI Table 3.1 – Metadata for each of the 51 environmental variables input to the logistic regression gene models, including variable name, time span metric over which historical sources are averaged or sample/satellite sources are gathered, original units of the dataset, whether data is available a depth resolution or surface only, and the original data source.

SI Table 3.2 – Satellite and historical environmental data corresponding to the sampling site, date, and depth of each TARA site.

SI Table 3.3 – The cluster centroid coefficient for each environmental feature. Columns: 'feature', environmental feature name; cluster1-5: centroid coefficient for that cluster.

SI Table 3.4 – Cluster membership proportions and Shannon Diversity metric for each TARA site. Columns: 'run_id', the European Nucleotide Archive run accession ID; 'longitude' and 'latitude': coordinates at which TARA site was sampled. 'tara_label': TARA project label. 'cluster1-5': proportion of genes present at that site assigned to that cluster. 'shannon': shannon diversity metric for each TARA site.

SI Table 3.5 – Average nucleotide identity pairwise comparisons of SAR86A-E. Each cell is the ANI value for genome 1 (rows) compared against genome 2 (columns). ANI calculated after (Varghese et al. 2015) using the online calculator at https://ani.jgi-psf.org/html/calc.php?.

REFERENCES

Antonov JI, Seidov D, Boyer TP, Locarnini RA, Mishonov AV, Garcia HE, et al. (2010). World Ocean Atlas 2009, Volume 2: Salinity. U.S. Government Printing Office.

Garcia HE, Locarnini RA, Boyer TP, Antonov JI, Baranova OK, Zweng MM et al. (2010). World Ocean Atlas 2009, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation. U.S. Government Printing Office.

Garcia HE, Locarnini RA, Boyer TP, Antonov JI, Baranova OK, Zweng MM et al. (2010). World Ocean Atlas 2009, Volume 4: Nutrients (phosphate, nitrate, silicate). U.S. Government Printing Office.

Hoarfrost A. SAR86. *Github repository*. https://github.com/ahoarfrost/SAR86/

Jickells TD, An ZS, Andersen KK, Baker AR, Bergametti G, Brooks N et al. (2005). Global iron connections between desert dust, ocean biogeochemistry, and climate. *Science* **308**:67–71.

Locarnini RA, Mishonov AV, Antonov JI, Boyer TP, Garcia HE, Baranova OK, et al. (2010). World Ocean Atlas 2009, Volume 1: Temperature. U.S. Government Printing Office.

Montegut CB, Madec G, Fischer AS, Lazar A. (2004). Mixed layer depth over the global ocean: An examination of profile data and a profile-based climatology. *Journal of Geophysical Research* **109**:C12003.

NASA earth observations. (2012). http://neo.sci.gsfc.nasa.gov/.

NASA Goddard Space Flight Center, Ocean Ecology Laboratory, Ocean Biology Processing Group; (2014): Sea-viewing Wide Field-of-view Sensor (SeaWiFS) Ocean Color Data, NASA OB.DAAC. http://doi.org/10.5067/ORBVIEW-2/SEAWIFS_OC.2014.0.

Ocean Productivity Group, Standard VGPM Model; http://www.science.oregonstate.edu/ocean.productivity/index.php

Ready J, Kaschner K, South AB, Eastwood PD, Rees T, Rius J et al. (2010). Predicting the distributions of marine organisms at the global scale. *Ecol Model* **221**:467–478.

Stott J. (2012). Earthtools. http://www.earthtools.org/webservices.htm.

Tyberghein L, Verbruggen H, Pauly K, Troupin C, Mineur F, Clerck OD. (2012). Bio-oracle: a global environmental dataset for marine species distribution modelling. *Global Ecol Biogeogr* **21**:272–281.

Varghese NJ, Mukherjee S, Ivanova N, Konstantinidis KT, Mavrommatis K, Kyrpides NC, et al. (2015) Microbial species delineation using whole genome sequences. *Nucleic Acids Res*; **43**: 6761–6771.

Frank Wentz, Simon Yueh, Gary Lagerloef. 2014. Aquarius Level 3 Sea Surface Salinity Standard Mapped Image Annual Data V3.0. Ver. 3.0. PO.DAAC, CA, USA.

APPENDIX D: CHAPTER 4 METASEEK EXPLAINER

The MetaSeek data discovery tool facilitates the search and discovery of sequencing datasets, integration of curated datasets along any of their available metadata, and the download and dissemination of these curated datasets. The publication associated with the MetaSeek tool was submitted to the *Application Notes* section of the *Bioinformatics* journal, which has strict length limits. Since some readers may be interested in a more detailed explanation of key aspects of the MetaSeek tool, this section will elaborate on the central components MetaSeek.

The basic workflow for MetaSeek starts by ingesting new metadata for sequencing datasets as they become publicly available, cleaning, parsing, and predicting missing metadata as it comes in. These cleaned metadata, as well as the original values where applicable, are stored in the MetaSeek database. An API communicates with this database, translating requests for information from the website or API users to retrieve the appropriate response from the database, and returning the results of the request to the user (either the website itself or a developer using the API for their own purposes). For any particular type of data retrieval request, there needs to be an associated API call programmed into the MetaSeek app, which determines both what kind of request information it expects to receive from the user, and the nature of the response from the database. The MetaSeek front end website receives the results from an API call, and displays them accordingly, which may vary depending on which web page is making the request and what part of the MetaSeek interface is being interacted with.

The following sections describe details of each piece of this workflow. Additional information, tutorials, and documents can also be found on the MetaSeek website (https://www.metaseek.cloud/) and Github repository (https://github.com/MetaSeek-Sequencing-Data-Discovery/metaseek).

Metadata scraping, cleaning, and parsing

MetaSeek hosts metadata from *all* sequencing datasets that are publicly available in the Sequence Read Archive (SRA), which includes a wide variety of sequencing data types. The only unifying quality of these datasets is that they are produced by next-generation sequencing technologies, and that they are publicly available (the SRA does allow a temporary privacy embargo for yet-to-be-published datasets). Periodically, MetaSeek runs a series of scraper scripts that are designed to find new datasets in the SRA, parse and clean them, and add them to the MetaSeek database. The MetaSeek scrapers use the NCBI eutilities, the NCBI's own API, to search for datasets and get metadata for each dataset. Complicating the issue is that the NCBI hosts a number of databases, all designed to store slightly different information, and the metadata for one dataset may be spread across multiple databases: not just the accession associated with the dataset's SRX accession ID (the ID associated with the SRA itself), but also additional sample metadata that may exist in the BioSample database and any publication metadata in PubMed.

In the first step of the scraper scripts, the MetaSeek scrapers use the eutilities "esearch" call to search for all publicly available datasets. Once it has a list of all dataset unique IDs in the SRA, it compares this list to the db_source_uid in the MetaSeek database, and removes any datasets that already exist. This leaves the scraper with a list of new SRA unique IDs for which metadata can be retrieved and added to MetaSeek. The SRA defines unique datasets at the level of the experiment, or SRX number, but these "experiments" can be nested within a larger project (SRP), and encompass multiple samples (SRS), or multiple sequencing runs (SRR). In MetaSeek, to avoid this nested structure, each row in the main MetaSeek Dataset table corresponds with an SRX number. If multiple runs are associated with a single SRX, there is an additional Runs table that stores just the relevant sequencing run information for each individual run.

In batches of 500 datasets at a time, the scrapers then scrape metadata from the SRA database itself using the eutilities “efetch” call. For each unique SRA accession, the efetch API call returns a structured XML file that contains much of the metadata MetaSeek needs. This metadata, as the SRA organizes it, falls into seven basic categories: 'EXPERIMENT',' SUBMISSION', 'Organization', 'STUDY', 'SAMPLE', 'Pool', and 'RUN_SET'. The Experiment section includes basic info like the SRX accession, experiment title, and study title, but also several metadata fields describing the basic nature of the type of sequencing data of the dataset such as library_source, library_strategy, library_screening_strategy, and library_construction_method, which correspond to parameters such as whether it is a genomic, metagenomic, or transcriptomic sample; whether whole genome sequencing or an amplification strategy was used; whether any sequencing selection such as PCR or restriction digest was used; and what sequencing platform was used to sequence the sample. These fields are extremely useful for MetaSeek because they are mandatory fields that all users must fill out when submitting metadata to the SRA (unlike most fields, including those that are mandatory for the MIxS metadata standards), there is only a small controlled vocabulary of value inputs that can be provided for each field, and they actually enforce this standard of entry when metadata is being submitted with a sequencing data submission so the metadata in the SRA database is very clean. A “controlled vocabulary” is the set of possible values that are supposed to correspond to a particular metadata field, as defined by a metadata standard group such as MIxS. The ‘Submission’ section includes the SRA submission ID. The ‘Organization’ section describes contact and location info for the submitter. The ‘Study’ section describes information about a larger study that may encompass the experiment, such as the study title and the study accession SRP id. The ‘Sample’ section includes some of the information in BioSample, if it exists, such as the sample id (SRS#), BioSample id (SAMN#), sample title and description, and information on the taxon scientific and common name, and NCBI taxon id, which describe the species or taxon if it’s a genomic sample. The ‘Pool’ section is redundant with the other sections and is skipped. The ‘Run’ section describes information about the actual sequencing run that was conducted: the number of reads and bases sequenced, the download size of the sequencing data, counts of the number of each nucleotide, and read quality counts. From this information MetaSeek also calculates the average read length and the GC percent.

Once the SRA metadata is collected, the scrapers use the eutilities “elink” call to identify whether any BioSample or PubMed accessions linked to the SRA dataset exist. If they do, metadata from these databases are gathered with an “efetch” command and added to the MetaSeek metadata for that dataset. The BioSample database, in theory, stores the MIxS standard metadata fields for a dataset, and users are encouraged to submit metadata that is compliant with these standards. From the BioSample metadata, in addition to basic sample title, description info, and collection date information, MetaSeek collects the biosample_package field that theoretically corresponds with the env_package field defined in the MIxS standards (whether a sample is from ‘sediment’, ‘human-gut’, etc.), as well as a mishmash of “sample_attributes” fields. These are free text fields provided by the user, that are supposed to be the MIxS-compliant mandatory and optional fields. A user enters both the field name and the field value as free text in their metadata submission packet, and these fields and values are supposed to correspond to MIxS-defined fields and controlled vocabulary values for the type of sample they are submitting, which the user is left to determine from a long set of guides and lists provided by the SRA in an excel worksheet. As one would expect, this metadata is extremely messy, riddled with misspellings and missing information, but potentially useful. The PubMed metadata, if it exists, is also collected via an “efetch” call, and this metadata includes the citation information for the publication, and the metadata publication date.

Once all of the raw metadata is collected from SRA, BioSample, and PubMed databases, MetaSeek applies a number of cleaning and parsing operations to try to parse the messy sample_attributes fields into a clean set of MIxS-standard compliant field names and the appropriate controlled vocabulary value. First, the user-provided field names are parsed to MIxS-compliant field names for all of the mandatory MIxS fields and the most commonly provided optional fields. This is done from a set of manually curated rules: for example, if a user has entered any of the values "env package", "environment package", "environmental package", "enviornmental package", "water environmental package", "human gut environmental package", etc., convert this field name to the MIxS-compliant “env_package”. These rules were curated using the first million datasets downloaded into the MetaSeek database during the app’s development, looking within the “sample_attributes” entries for the most common misspellings and misused terms, and parsing these to MIxS-compliant fields. When all of the MIxS-compliant fields are parsed and extracted, the user-provided values for these fields are parsed to comply with the appropriate controlled vocabulary for those fields that require them. This is done for the “investigation_type”, “env_package”, “sequencing_method”, and “mixs_specification” fields. Again, this is done using a set of rules – “if you see this or this or this change it to this” – that were manually curated by inspecting the most common mistakes and errors in one million user-provided metadata entries.

One field, “investigation_type”, is often missing metadata entries entirely despite the fact that it is a key (and theoretically “mandatory”) MIxS field, and can’t be parsed manually. However, this field can be accurately predicted from the mandatory SRA fields that are enforced during the SRA metadata collection process, are available for every dataset, and have clean controlled vocabularies. In the case where “investigation_type” is missing or can’t be parsed manually, I use a logistic regression model to predict the “investigation_type” field from the “library_source”, “library_strategy”, “library_screening_strategy”, and “study_type” mandatory SRA metadata fields. This can predict “investigation_type” with 94.1% accuracy. A confidence field with a measure of the confidence of the model prediction, “metaseek_investigation_type_P”, is also recorded. In all cases where MetaSeek has changed an original user entry to produce clean metadata, whether manually parsed or predicted, the original user entry is maintained in the “sample_attributes” field.

Two other important fields that is very messy are the latitude and longitude fields. The values for latitude and longitude are provided by the user as free text, which predictably results in many entries that are not easily converted to numeric values, and may use different units such as degrees-minutes-seconds rather than decimal degrees. However, there are a small number of common entry formats that users tend to use, and this can capture the vast majority of latitude/longitude entries. Using a series of regular expression pattern matching, these patterns are recognized and converted to a numeric value in decimal degrees, which is recorded in the “meta_latitude” and “meta_longitude” fields. The original “latitude”, “longitude”, or “lat_lon” fields are maintained in the MetaSeek database as well.

In a final cleaning step, missing data values which are often recorded by a variety of values – 'NA', 'Missing', 'missing', 'unspecified', 'not available', 'not given', 'Not available', 'not applicable', etc. – are converted to a uniform None value. Finally, the cleaned and parsed MetaSeek metadata is inserted into the MetaSeek database.

The MetaSeek database and API

The bulk of the MetaSeek database, containing the scraped and cleaned metadata, is contained in a single table, the Dataset table. Each row corresponds to a single SRX accession. The MetaSeek database also contains a Run table, which records sequencing information such as number of reads and bases sequenced about the SRR accessions associated with an SRX, and a Publication table, which records the PubMed publication metadata, if it exists, associated with a particular SRX. A Discovery table records the saved filter parameters, timestamp, owner, and "discovery" title and description for any saved discoveries from the MetaSeek website, and a User table saves the user ID associated with any users who log in on the MetaSeek website with the "log in with Google" button. These tables are used to retrieve discoveries saved by a user at a later date.

This database is hosted by Redshift, which is an Amazon database service that streamlines much of the tedious database management tasks, and is also uses a columnar data storage schema. This is in contrast to MySQL, for example, which uses a row-wise data storage schema. Databases, in general, are optimized to do particular operations lightning fast, because they organize the data under the hood in such a way that retrieving and returning certain kinds of information is optimized. However, this comes with tradeoffs, where other kinds of operations are not optimized and slower. In a traditional row-wise database, grabbing a row from the database and returning it to the user is *super* fast, but doing column-wise operations like filtering on the values of multiple fields can be slow. A columnar data storage is optimized in the opposite direction, such that retrieving summary data from several columns is optimized for. This columnar optimization is ideal for the MetaSeek web interface: in the main MetaSeek "Explore" page, every time a user sets a filter parameter, they are asking the MetaSeek database to filter a field in a column-wise operation. Especially as databases get big, and MetaSeek now has over 3 million unique entries in the Dataset table, this can become very slow, and as a result for the MetaSeek Explore page this can mean you wait a long time to get search results back. MetaSeek was originally stored in a MySQL database, and was recently converted to a columnar Redshift database. During performance testing of the new schema, queries that took 30 seconds to 1 minute to return results in the MySQL database took between 0.5-5 seconds to return results in the Redshift database.

The MetaSeek API provides a set of calls designed to retrieve information required by the website interface, and also provides a number of user-facing API calls for programmatic access to query the MetaSeek database. This API is written in the python Flask library, which interfaces well with the web as well as multiple database schemas. The database-Flask app-React website software stack is a common use case, which has the added advantage of having a lot of documentation for overcoming common stumbling blocks on the web. The API receives requests as they come in – either from the MetaSeek website, or from a user accessing the database programmatically – and translates these requests to the database, which returns the appropriate information. The code of the Flask app defines the set of API calls that can be received, how that request should be parsed, and the kind of result that should be returned. For example, when a user selects a filter on the MetaSeek "Explore" page, the website sends to the app the filter parameters the user has set: the field filtered, the value/s the field should contain, and how this field should be filtered, e.g. values less than, or greater than, or included in a list of acceptable values, etc. The MetaSeek API takes this information and translates it into a NoSQL query that the Redshift database can understand, collects the results, and packages the returned data into the summary data visualized in the MetaSeek Explore visualization dashboard. Other API calls are available as well. When a user queries the API programmatically with the "SearchDatasetIds", API call, for example, they similarly provide filter parameter data, which the Flask app translates into the appropriate NoSQL query, the database returns the MetaSeek dataset IDs that match those filter parameters, and the API packages this into a json format and sends this back to the user.

The MetaSeek web interface

The meat of the MetaSeek website is on the Explore page. This is where users can filter the MetaSeek database based on the most important metadata fields, and look at interactive visualizations that summarize the matching datasets in real time. The most important fields are broken down into "General", "Sequencing", and "Environmental/Contextual" info. The general info includes basic information such as what type of sequencing dataset it is – a genome, metagenome, marker gene survey, etc. – and the basic source of the dataset – whether it's from sediment, soil, water, human gut, wastewater, etc. The sequencing info section provides filters for the sequencing platform used (Illumina, Nanopore, etc.), the average read length or number of reads sequenced, whether it was single or paired-end sequencing, and the library strategy or screening strategy on which you can filter for WGS vs. amplicon, or Random vs. PCR sequencing. The environmental section provides filters for latitude, longitude, and the environmental biome, feature, material, and geographic location fields (MIxS fields describing the general environmental context of the sample). When a filter parameter is set, this is sent to the API, which retrieves the datasets that match the filter parameters, summarize the results, and send this summary data back to the browser. The Javascript code on the website then uses this summary data to produce a suite of interactive visualizations on the visualization dashboard that allows the user to explore the features of the results of their search. These visualizations generally visualize the counts of datasets that fit in certain categories, or display histograms of the distribution of values from numeric fields across all the datasets, or plot the density of samples collected from a range of latitudes and longitudes on a map. There is also a paginated table that allows users to look at the full dataset metadata for matching datasets; by paginating this table, the browser doesn't need to store the full metadata of *all* of the matching datasets, but only a few at a time, allowing a user to theoretically look at each dataset manually without crashing the browser.

If a user finds a set of filter parameters that match the datasets they are looking for and they want to save their results for a later date, or download the metadata or matching datasets to look at offline, they can save their "discovery" on the MetaSeek website. The site prompts the user to sign in with Google, and the discovery filter parameters are saved in the Discovery table of the MetaSeek database and associated with that user id in the User table of the MetaSeek database. The user is then taken to the "Discovery Details" page, which displays the visualization dashboard of the discovery's matching datasets, and provides links to download either the matching dataset IDs or the full metadata for that discovery. These discoveries can also be browsed by any user on the Browse page, and that metadata can be downloaded. By saving the filter parameters and timestamp of each discovery, a user is able to easily return to and retrieve results at a later date, or cite their discovery in a future publication.