

**Untangling interactions between uncultured marine microbes  
across domains of life through single-cell genomics and  
timeseries analysis**

Dissertation

vorgelegt von

Fabian Wittmers

Kiel, August 2024



**Untangling interactions between uncultured marine microbes  
across domains of life through single-cell genomics and  
timeseries analysis**

---

Dissertation

vorgelegt von

Fabian Wittmers

zur Erlangung des akademischen Grades eines Doktors der Naturwissenschaften

**Dr. rer. nat.**

an der Mathematisch-Naturwissenschaftlichen Fakultät der  
Christian-Albrechts-Universität zu Kiel

Kiel, August 2024

**1. Gutachterin: Prof. Dr. Alexandra Worden**

**2. Gutachter: Prof. Dr. Eric Achterberg**

**Tag der Disputation: 15. November 2024**

# CONTENTS || INHALTSVERZEICHNIS

|                                                                                         |            |
|-----------------------------------------------------------------------------------------|------------|
| <b>Summary</b>                                                                          | <b>I</b>   |
| <b>Zusammenfassung</b>                                                                  | <b>III</b> |
| <b>Acknowledgements</b>                                                                 | <b>V</b>   |
| <b>Eidesstattliche Erklärung</b>                                                        | <b>IX</b>  |
| <b>Chapter 1: General Introduction</b>                                                  | <b>1</b>   |
| Microbial life on a microscale                                                          | 1          |
| Protists in long-term timeseries                                                        | 4          |
| Thesis Outline                                                                          | 4          |
| References                                                                              | 7          |
| <b>Chapter 2</b>                                                                        | <b>9</b>   |
| Abstract                                                                                | 11         |
| Introduction                                                                            | 12         |
| Material & Methods                                                                      | 15         |
| Single cell sorting and amplification                                                   | 15         |
| Single-protist-well DNA sequence assemblies and ipaMAG recovery                         | 16         |
| 16S rRNA gene screening & phylogenetic placement                                        | 17         |
| Detection of symbiont-lineages in global ecosystems                                     | 18         |
| Functional profiling of recovered bacterial genomes and reference genomes               | 18         |
| Data and code availability                                                              | 19         |
| Results & Discussion                                                                    | 20         |
| Marine protists associated bacteria are dominated by five lineages                      | 20         |
| Variations in host range, genome size and GC content                                    | 24         |
| General functional trends                                                               | 27         |
| Bacteroidota, Verrucomicrobiota and Planctomycetota as members of the phycosphere       | 28         |
| Carbohydrate-Active EnZyme diversity within phycosphere-bacteria.                       | 31         |
| Motility systems and cell surface attachment mechanisms                                 | 36         |
| Global distribution and particle-attachment of bacteria associated with marine protists | 38         |
| Conclusions                                                                             | 40         |
| Author Contributions                                                                    | 43         |
| References                                                                              | 44         |

|                                                                                                                 |            |
|-----------------------------------------------------------------------------------------------------------------|------------|
| <b>Chapter 3</b>                                                                                                | <b>55</b>  |
| Abstract                                                                                                        | 56         |
| Introduction                                                                                                    | 57         |
| Material & Methods                                                                                              | 59         |
| Completeness and contamination refinement of high quality ipaMAG-0934                                           | 59         |
| 16S rRNA gene phylogenies                                                                                       | 59         |
| Phylogenetic reconstruction of horizontal gene transfer through gene trees                                      | 60         |
| Phylogenomics                                                                                                   | 61         |
| Ortholog analysis                                                                                               | 63         |
| Data availability                                                                                               | 63         |
| Results & Discussion                                                                                            | 64         |
| Discovery of globally distributed uncultivated marine bacterial lineages co-associated with predatory protists. | 64         |
| Reflections of novel marine bacterial-predatory protist co-associates on the evolution of animal pathogens.     | 65         |
| Nuances in mechanisms and metabolic approaches implicated in symbiosis.                                         | 74         |
| Trajectories of symbioses with marine predatory protists – independent routes.                                  | 76         |
| Gain and loss differentiate the Simplicirickettsiaceae and Rickettsiaceae                                       | 82         |
| Environmental importance and the optimal host                                                                   | 84         |
| Conclusions                                                                                                     | 89         |
| Author Contributions                                                                                            | 90         |
| References                                                                                                      | 91         |
| Supplementary Figures                                                                                           | 100        |
| <b>Chapter 4</b>                                                                                                | <b>107</b> |
| Abstract                                                                                                        | 108        |
| Introduction                                                                                                    | 109        |
| Material & Methods                                                                                              | 110        |
| Single cell sorting, amplification, and sequencing                                                              | 110        |
| Recovery of Margulisbacteria assemblies                                                                         | 110        |
| Phylogenomics                                                                                                   | 112        |
| 16S rRNA gene phylogeny                                                                                         | 113        |
| V4 16S rRNA gene amplicon analysis                                                                              | 114        |
| Data Availability                                                                                               | 114        |
| Results & Discussion                                                                                            | 115        |
| Conclusions                                                                                                     | 119        |
| Acknowledgements                                                                                                | 120        |
| Author Contributions                                                                                            | 120        |
| References                                                                                                      | 121        |

|                                                                                                                             |            |
|-----------------------------------------------------------------------------------------------------------------------------|------------|
| <b>Chapter 5</b>                                                                                                            | <b>125</b> |
| Abstract                                                                                                                    | 126        |
| Introduction                                                                                                                | 127        |
| Material & Methods                                                                                                          | 129        |
| Oceanographic sampling                                                                                                      | 129        |
| Physical framework                                                                                                          | 129        |
| DNA extraction and sequencing                                                                                               | 130        |
| Phylogenetic classification of amplicons                                                                                    | 130        |
| Diversity metrics                                                                                                           | 131        |
| Co-occurrence network analysis                                                                                              | 131        |
| Results & Discussion                                                                                                        | 134        |
| Species turnover and change in community composition over time                                                              | 135        |
| Relative decline in surface phytoplankton                                                                                   | 138        |
| Changes in <i>Prochlorococcus</i> ecotype composition                                                                       | 140        |
| Changes in eukaryotic surface phytoplankton - the rise of the chrysophytes                                                  | 142        |
| Dynamics in other putatively mixotrophic lineages                                                                           | 146        |
| Changes in the abundance of highly interactive eukaryotic phytoplankton over time causes changes in heterotrophic community | 152        |
| General limitations of network analysis                                                                                     | 153        |
| Conclusions                                                                                                                 | 154        |
| Author Contributions                                                                                                        | 155        |
| References                                                                                                                  | 156        |
| Supplementary Figures                                                                                                       | 163        |
| <b>Chapter 6</b>                                                                                                            | <b>169</b> |
| Broad importance of newly discovered marine protist-associated symbiont lineages                                            | 170        |
| Challenges in eukaryotic single-cell genomics                                                                               | 171        |
| Co-cultivation of protists is a necessary complement to culture-independent approaches                                      | 173        |
| Scales – implications of single-cell interactions for global pathways and the marketplace of chemical currencies            | 173        |
| Cell - to - cell interactions in co-occurrence network models                                                               | 174        |
| Long-term timeseries                                                                                                        | 175        |
| References                                                                                                                  | 177        |

## Summary

Microbes of all lifeforms build the foundation of the marine food web and base of the carbon cycle. Diverse unicellular eukaryotes, many uncultivated, occupy key positions in the food web and have evolved complex lifestyles and interact extensively with bacteria in the surface ocean, with major implications for the flow of energy in the ecosystem. Beyond predator-prey relationships, both photosynthetic and phagocytotic protists engage in relationships with other microbes going beyond classical predator-prey interactions, but those remain poorly characterized. This microbial community and their interactions across lifestyles will likely change as a consequence of natural and anthropogenic forcing over time, driving changes in temperature, stratification and nutrient inventory, among other environmental parameters. We utilized single cell sorting and whole genome sequencing of marine protists to discover bacteria associated with photosynthetic and heterotrophic protists. We then analysed microbial taxa at the Bermuda Atlantic Timeseries Study (BATS) site over the past 30 years using V1V2 16S rRNA gene amplicon sequencing to reveal changes in the composition of surface phytoplankton and interactions with the heterotrophic community.

We observed a clear separation in taxonomy and genome characteristics for symbionts of protists with different lifestyles. Putative symbionts from the Alphaproteobacteria and Gammaproteobacteria associated with heterotrophic protists included four new lineages with small genome sizes related to either the Rickettsiales or the Coxiellales, orders which include many intracellular symbionts and pathogens of metazoans. They resided within choanoflagellates and the MAST-3 lineages, dominant consumers of the most abundant phytoplankton and other microbes. Additionally, three lineages of Marinamargulisbacteria, a group of elusive bacteria sharing a progenitor with the Cyanobacteria, were found in association with Choanoflagellates and multiple heterotrophic marine stramenopile lineages (MAST-4 and MAST-1), providing the first indication that the Marinamargulisbacteria are symbionts of protists. This also constitutes the first known symbionts associated with any MAST lineage. The choanoflagellate and MAST-3 associated Alpha- and Gammaproteobacteria lineages, related to metazoan pathogens, including causative agents of disease in human, possess specialized adaptations in the host environment, including complete Type IV secretion systems. However, their manner of symbioses differs, following distinct trajectories. A lineage of Alphaproteobacteria (Simplirickettsiaceae) showed extensive genome reduction to a genome size <1 Mb, losing key metabolic functionality suggesting an obligate symbiosis with their choanoflagellate hosts. An order of

Gammaproteobacteria, previously not known to be host associated (Protistiellales) showed signs of extensive horizontal acquisition of genes from diverse, unrelated bacterial pathogens, and a lifestyle like *Legionella pneumonia*, preferentially associating with hosts but with some free-living capacity. The microbes associated with photosynthetic protists were primarily Planctomycetota, Verrucomicrobiota and Bacteroidota, with larger genomes than those of heterotrophic protist-associated bacteria and a diverse repertoire of non-essential metabolic functions, most significantly carbohydrate-active enzymes (CAZys). These CAZy pathways included enzymes suggesting the possibility of the cells attacking the cell wall itself, not purely relying on carbohydrates and dissolved organic matter secreted by the phytoplankton as by-products of photosynthesis. These members of the phycosphere surrounding the phytoplankton cell encoded pili and flagella, absent in the symbionts of heterotrophic protists, enabling varying degrees of motility and attachment in line with an extracellular lifestyle involving physical attachment to the host eukaryote. The diversity and evolutionary position of the novel symbiont of marine protists detected herein, could enable a better understanding of the evolution of bacterial pathogens of metazoans, as well as the ecology and complex resource flow in the phycosphere and beyond, observed within complex ocean food webs. Observations of microbial interactions changing over time indicates that these interactions are severely impacted by changes in environmental conditions, such as warming of surface waters and a decline in nutrient concentrations, especially in stratified summer conditions in the open ocean, through a decline in phytoplankton and subsequent impact on the heterotrophic community.

## Zusammenfassung

Mikroben aller Lebensformen bilden die Grundlage des marinen Nahrungsnetzes und die Basis des Kohlenstoffkreislaufs. Diverse einzellige Eukaryoten, von denen viele nicht durch Laborkulturen repräsentiert sind, nehmen Schlüsselpositionen im Nahrungsnetz ein, haben komplexe Lebensweisen entwickelt und interagieren ausgiebig mit Bakterien was erhebliche Auswirkungen auf den Energiefluss im Ökosystem hat. Sowohl photosynthetische als auch phagozytotische Protisten interagieren mit anderen Mikroben, die über die klassischen Räuber-Beute-Beziehungen hinausgehen, die jedoch nur unzureichend beschrieben sind. Diese mikrobielle Gemeinschaft und ihre Interaktionen über verschiedenen Lebensformen hinweg werden sich wahrscheinlich im Laufe der Zeit infolge natürlicher und anthropogener Einflüsse verändern, die neben anderen Umweltparametern auch zu Veränderungen bei Temperatur, Schichtung und Nährstoffinventar führen. Wir haben individuelle Zellen von marinen Protisten sortiert und deren Genome sequenziert, um Bakterien zu entdecken, die mit photosynthetischen und heterotrophen Protisten assoziiert sind. Anschließend analysierten wir die mikrobiellen Taxa an der Bermuda Atlantic Timeseries Study (BATS) Station über die letzten 30 Jahre mit Hilfe der V1V2 16S rRNA-Gene Sequenzierung analysiert, um Veränderungen in der Zusammensetzung des Oberflächenphytoplanktons und Interaktionen mit der heterotrophen Gemeinschaft aufzudecken.

Wir beobachteten eine klare Trennung in der Taxonomie und den genetischen Charakteristika für Symbionten von Protisten mit unterschiedlichen Lebensweisen. Zu den mutmaßlichen Symbionten der Alphaproteobakterien und Gammaproteobakterien, die mit heterotrophen Protisten assoziiert sind, gehörten vier neue Gruppen mit kleinen Genomen, die entweder mit den Rickettsiales oder den Coxiellales verwandt sind; Ordnungen, die viele intrazelluläre Symbionten und Krankheitserreger von Metazoen umfassen. Sie gehören zu den Choanoflagellaten und den MAST-3-Linien, den wichtigsten Prädatoren von Phytoplankton und anderer Mikroben. Darüber hinaus wurden drei Gruppen von Marinamargulisbakterien, einer Gruppe kaum charakterisierter Bakterien, die einen Vorläufer mit den Cyanobakterien haben, in Verbindung mit Choanoflagellaten und mehreren heterotrophen marinen Stramenopile-Linien (MAST-4 und MAST-1) gefunden. Dies liefert den ersten Hinweis darauf, dass die Marinamargulisbakterien Symbionten von Protisten sind. Dies sind auch die ersten bekannten Symbionten, die mit einer MAST-Linie assoziiert sind. Die mit Choanoflagellaten und MAST-3 assoziierten Alpha- und Gammaproteobakterien, die mit Krankheitserregern von Metazoen verwandt sind, einschließlich der Erreger von

Krankheiten beim Menschen, verfügen über spezialisierte Anpassungen an die Umgebung ihres Wirtes, einschließlich vollständiger Typ-IV-Sekretionssysteme. Die Art und Weise ihrer Symbiose ist jedoch unterschiedlich und folgt verschiedenen Entwicklungspfaden. Eine Linie von Alphaproteobakterien (Simplirickettsiaceae) zeigte eine umfassende Genomreduktion auf eine Genomgröße  $<1$  MB und gleichzeitigem Verlust wichtiger Stoffwechselfunktionen, was auf eine obligate Symbiose mit ihren Choanoflagellaten-Wirten schließen lässt. Eine Ordnung von Gammaproteobakterien, von der bisher nicht bekannt war, dass sie mit dem Wirt assoziiert ist (Protistiellales), zeigte Anzeichen für einen umfangreichen horizontalen Erwerb von Genen aus verschiedenen, nicht verwandten bakteriellen Krankheitserregern und eine Lebensweise wie bei *Legionella pneumonia*, wobei sie bevorzugt mit dem Wirt assoziiert sind, aber auch eine gewisse Fähigkeit zum freien Leben haben. Bei den mit photosynthetischen Protisten assoziierten Mikroben handelte es sich in erster Linie um Planctomycetota, Verrucomicrobiota und Bacteroidota mit größeren Genomen als bei den mit heterotrophen Protisten assoziierten Bakterien und einem vielfältigen Repertoire an nicht-essentiellen Stoffwechselfunktionen, vor allem Kohlenhydrate-aktiven Enzymen (CAZys). Diese CAZy-Stoffwechselwege enthielten Enzyme, die darauf schließen lassen, dass die Zellen die Zellwand selbst angreifen und sich nicht nur auf Kohlenhydrate und gelöste organische Stoffe stützen, die vom Phytoplankton als Nebenprodukte der Photosynthese abgesondert werden. Diese Mitglieder der Phycosphäre, die die Phytoplankton Zelle umgibt, kodierte Pili und Flagellen, die bei den Symbionten der heterotrophen Protisten fehlen, und ermöglichten so ein unterschiedliches Maß an Motilität und Anhaftung im Einklang mit einer extrazellulären Lebensweise, die eine physische Anhaftung an das Phytoplankton beinhaltet. Die Vielfalt und die evolutionären Positionen der hier entdeckten neuartigen Symbionten mariner Protisten könnte ein besseres Verständnis der Evolution bakterieller Krankheitserreger von Metazoen sowie der Ökologie und des komplexen Ressourcenflusses in der Phycosphäre und darüber hinaus ermöglichen, die in komplexen Nahrungsnetzen im Ozean zu beobachten sind. Beobachtungen der mikrobiellen Interaktionen, die sich im Laufe der Zeit verändern, deuten darauf hin, dass diese Interaktionen durch veränderte Umweltbedingungen, wie z. B. die Erwärmung des Oberflächenwassers und den Rückgang der Nährstoffkonzentrationen, insbesondere während des Sommers im offenen Ozean, durch einen Rückgang des Phytoplanktons und die daraus resultierenden Auswirkungen auf die heterotrophe Gemeinschaft stark beeinträchtigt werden.

## Acknowledgements

Before anyone else, I want to express my gratitude to my mentor and supervisor Alexandra Worden for guidance throughout this long journey, thoughtful discussions along the way and for keeping up with me being the stubborn & opinionated person I am. I have learned how to think about science and the big picture through many intriguing conversations, and you have given me the opportunity to immerse myself in the work that I wanted to do in my own way in the middle of a global pandemic, directing me when necessary but letting me find my own way through the maze of microbial ecology. I also want to thank Stephen J. Giovannoni for continuous availability and mentorship, thoughtful feedback, great discussions, and even better ideas on how I can improve my work and can become a better scientist; I have deep admiration about how you think about the microbial ocean and come up with fascinating new hypothesis and ideas, driving the field forward again and again, and hope to learn more from you in the future.

I want to thank the whole BIOS-SCOPE team for creating a project and collaborative environment full of great people. Science can be a cutthroat with more politics than actual science behind the scenes, but the amazing scientists I got to work with in this project have been welcoming, shared their insights, thoughts, and data openly with a naive doctoral student. I hope that I can create and maintain and keep working in an environment like that in the future. Rachel J. Parsons at the Bermuda Institute of Ocean Sciences and her team have been very good to me whenever I was on Bermuda for my doctoral work and cruises. I would not have been able to do the research cruises without their fantastic support. I appreciate the support of the Simons Foundation International, the funding body of the BIOS-SCOPE project, for giving me the opportunity to conduct my doctoral work within this group of great scientists and collaborators.

I owe a to the scientists that created and, for decades, collected data and samples at the Bermuda Atlantic Timeseries Study (BATS) site. A special thanks to Stephen J. Giovannoni and Craig A. Carlson and their teams who have collected DNA at BATS starting in 1991, years before I was even born. I relied heavily on meticulously curated metadata available for BATS, generated and maintained by the BATS team at BIOS, currently led by Rod Johnston and Nicolas Bates. Dozens of scientists have contributed to this tremendous undertaking for decades, too many to name them all, but the contribution of each of them made this timeseries possible. Through their long-term vision and effort, I was able to utilize this unique dataset during my dissertation. Thanks to Shuting Liu, Jacqui Comstock, Michelle

Michelsen and Kevin Vergin for spending weeks in the lab with me preparing sequencing plates for thousands of samples that I got to use for my final chapter. I also owe a lot to the long-time Worden Lab support scientist Camille Poirier, whose phenomenal skills in single cell sorting and sequencing helped tremendously in generating much of the raw data used throughout this thesis.

Throughout my time as a doctoral student, I was able to participate in multiple research cruises, meeting fantastic scientists, many of whom became friends. Field work can be demanding and hectic, but many collaborators, especially within the BIOS-SCOPE project volunteered their expertise and little off-time to teach me all their tips and tricks. I am particularly thankful to Krista Longnecker, Elizabeth Kujawinski, Kristin Bergauer, and Craig A. Carlson who has been an impressive, dedicated, and hands-on chief-scientist for five cruises I participated in. I appreciate the captain and crew of the R/V *Western Flyer*, R/V *Pourquoi Pas*, the R/V *Sonne* and especially the R/V *Atlantic Explorer* for their hard work, making all the fieldwork possible. A special thanks to Dexter and his kitchen-team for providing phenomenal food and fresh cookies daily on the R/V *Atlantic Explorer*, it was at times the only thing keeping me alive after days of barely any sleep.

I want to express my gratitude to all members of the Worden Lab, both at the GEOMAR Helmholtz Centre for Ocean Research and the Marine Biological Laboratory in Woods Hole. I need to highlight Yu-Chen Ling who I shared an office with for years for her endless supply of chocolate at times when I was thinking about throwing my laptop out of the window. Rachele Spezzano, Charlotte Eckmann and Maria Hamilton have been great colleagues and friends along the way. I learned how to be a better scientist and person from everyone, but especially Olga Matantseva and Kristin Bergauer who provided me valuable insights and taught me a lot personally, in and outside of the lab, and in the field.

I would have probably quit my job a dozen times already if I didn't have the friends I have found in my fellow BIOS-SCOPE PhD students, Michelle Michelsen, Lillian Henderson, Chance English, and Noah Germolus. It takes a special group of folks to make demanding fieldwork with little to no sleep for days an experience that I looked forward to every time, because I got to spend time with some of my favorite people. I hope I get to keep working with you in the future; meeting you guys its probably the best thing that could have happened to me during my dissertation.

I also need to shoutout cycling, which has probably been the sport and passion outside of science that has kept me sane at time when everything else was overwhelming. Alongside my bike, my housemates, friends, and family have been a support system long before I even

## VII

embarked on the journey that was my dissertation. I have fond memories of great Sunday brunches in Kiel with excessive amounts of pancakes and Nutella with people I got to live with or do my Master's at GEOMAR with. My parents Meinolf and Susanne, especially my mother, have been my biggest supporters, and this dissertation is dedicated to her. Without her I would have not made it this far. She believed in me when no one else, including myself, did. I still remember vividly when my primary school teacher told me and my parents that I wasn't clever enough to make it to a high-school degree and send me off. My mother always saw something in me that my teacher and no one else saw and went to bat for me to give me a chance at higher education. Turns out she was right.



## Eidesstattliche Erklärung

Hiermit erkläre ich, Fabian Wittmers, dass die vorliegende Dissertation mit dem Titel

### **Untangling interactions between uncultured marine microbes across domains of life through single-cell genomics and timeseries analysis**

von mir selbstständig verfasst worden ist und nach Inhalt und Form meine eigene Arbeit ist. Es wurden keine weiteren Quellen und Hilfsmittel als die angegebenen verwendet. Ich habe die Arbeit unter Einhaltung der Regeln guter wissenschaftlicher Praxis der Deutschen Forschungsgemeinschaft angefertigt. Mir wurde kein akademischer Grad entzogen. Die vorliegende Arbeit wurde weder ganz oder zum Teil schon an anderer Stelle im Rahmen eines Prüfungsverfahrens vorgelegt. Veröffentlichte Kapitel oder solche, die bei einem Fachmagazin zur Veröffentlichung als Artikel eingereicht wurden, sind kenntlich gemacht.

## **Chapter 1: General Introduction**

Unicellular lifeforms are a cornerstone for the functionality of marine ecosystems, comprising bacteria, archaea, and single-cell eukaryotes (protists) involved in all biogeochemical processes taking place in the ocean [1]. These microbes have diverse lifestyles, ranging from heterotrophic bacteria that play an essential role in the remineralization of organic carbon as part of the microbial loop [2, 3] to phytoplankton, which, although contributing only 1% of global photosynthetic biomass, does account for roughly 50% of global primary productivity [4]. Primary producers including cyanobacteria and eukaryotic phytoplankton fix 50 Pg of atmospheric CO<sub>2</sub> yearly [4], building the basis of the open ocean food web and making them a cornerstone of the global carbon cycle. Heterotrophic protists, on the other hand, are feeding on other microbes, including phytoplankton, consuming and remineralizing carbon, but also moving carbon along the food-chain to higher trophic levels [1]. Both photosynthetic and heterotrophic protists are extremely diverse, often hard to quantify and mostly uncultured, leading to a lack of fundamental knowledge on many of these lineages [1].

### **Microbial life on a microscale**

Although of global importance, microbial life (and death) predominantly takes place on a microscale and ‘microbes see a sea of gradients’ highlighting the heterogeneity of the ocean on a small scale [5]. This heterogeneity is driven by gradients in the concentration of chemicals of diverse origin, for example, caused by the punctual release of organic molecules by photosynthetic protists and their subsequent diffusion, or components released by cells during cell death. What matters for the survival of a microbe is only what is present in its immediate microenvironment. For example, a microbe can grow rapidly if there is an abundance of easily accessible organic carbon present in its immediate environment that it can remineralize to produce energy, even when living in an open ocean environment, which is that contains only low concentrations of nutrients on a larger scale [5]. These hotspots can be created by phytoplankton and the materials these cells release during photosynthesis. Excess dissolved organic matter (DOM) is exported (passively and actively) to the outside environment, where it accumulates and slowly diffuses in the microscale environment surrounding the cell (estimated ~1,200 µm radius for a diatom with a diameter of 50 µm, but heavily dependent on the size of the phytoplankton cell) [6]. This organic matter is a major driver that attracts the phytoplankton cells microbiome, composed of heterotrophic bacteria

that respire DOM and in turn provide cofactors and inorganic macronutrients benefiting the phytoplankton cell. On the other hand, both heterotrophic bacteria and phytoplankton require the same inorganic nutrients (especially nitrogen, phosphate, and iron), leading to competition between them. Holistically, this makes the phycosphere a microenvironment of extensive microbial interactions on a small scale. These interactions include symbiotic relationships, defined as the living together between different species, spanning a spectrum of interactions ranging from mutualistic (benefiting both the host and symbiont) towards pathogenic (one part, the pathogen exploiting the other, the host, for its own benefit) [7]. Microbes living in the phycosphere can be described as the phytoplankton's microbiome and comprises the microbial community they harbour intracellularly or on their surface, similar to the microorganisms found on human skin or in the gut of all animals. Heterotrophic protists can also harbour a microbiome, e.g., when containing an intracellular pathogen or a microbe attached to its cell surface. The diversity of these symbiotic relationships in heterotrophic protists is likely just as diverse as symbiosis seen in other microbes, but its discovery is challenging.

Our current insights into symbionts of heterotrophic protists are predominantly based on the utilization of model organisms, mainly the ciliate genus *Paramecium* [8, 9] and amoebae [10, 11], lineages known to frequently engage in symbiotic relationships with various bacteria, which might use the ciliate as an opportunistic host. This, however, creates a biased insight into both host and symbionts. First, although this approach can help in identification of symbionts, it does miss all symbiotic lineages not adapted to ciliate hosts, including those that might be symbionts of key marine heterotrophic protists. Second, we do not get insights into the symbionts preferred host, their host range and host specificity. This is key because symbionts can have extremely different effects on different hosts, benefiting one while essentially representing a pathogen of the other at the same time, essentially shifting along the spectrum of symbiosis when moving from one host to another [12, 13]. Some of the key predatory protists in marine ecosystems are heterotrophic nanoflagellates, but a majority of those lineages have no or few cultured representatives [14, 15], leading to a lack of insights into their potential microbiome. Additionally, symbiotic relationships can change depending on environmental conditions, shown, e.g., in diatoms where an increase in temperature resulted in a Rhodobacterales bacterium in the phycosphere turning into a pathogen [16]. Considering major changes in the open ocean environment, driven by an increase in atmospheric CO<sub>2</sub> causing warming of surface water, less vertical mixing of the surface ocean and a subsequent decline in nutrient inventory, amongst other [17, 18], which poses the

question how microbial interactions will change in a changing ocean. Generally, more oligotrophic conditions would lead to a decrease in primary productivity with subsequently more dominance of recycling of organic matter by the microbial food web, predicted to lead to a decrease in carbon export [19].

Characterizing the microbiome of protists and their interactions with prokaryotic symbionts is intrinsically difficult. In culture, many phytoplankton isolates are kept axenic, in monocultures free of bacteria, essentially removing all the potential symbionts and members of the microbiome. Protists that rely on their microbiome for survival are categorically excluded in those cultivation efforts because they cannot survive in isolation. Many microbes, especially abundant heterotrophic protists, and key predators in the microbial foodweb have evaded cultivation-efforts for decades and their interactions with other microbes, including predator-prey relationships cannot be studied using any culture-based approaches [15]. The rise of large-scale amplicon-based and sequencing surveys in the last two decades have enabled an unprecedented insight into diversity and uncultured lineages of marine microbes and subsequent correlation-based networks have led to the prediction of countless potential biotic interactions between marine microbes [20, 21]. Similarly, metagenomics (sequencing of DNA generated from bulk filtration of seawater) and metatranscriptomics (similar to metagenomics but utilizing RNA transcriptomes) approaches have provided new insights into the functional potential of marine microbes [22, 23], including various pathways and genomic systems that point at symbiotic relationships and interactions with other microbes. However, these sequencing-based approaches do come with major limitations. Amplicon and meta-omics approaches are bulk sequencing techniques that rely on DNA recovered from large volumes of water and the community is essentially homogenised before sequences, making resolution of fine-scale heterogeneity in the water column impossible. Metagenomics is further challenged by biased assemblies, not recovering no eukaryotic genomes and no rare lineages nor particularly abundant populations of microbes characterized by heterogeneity in their genomes, which includes both the most abundant marine heterotrophic bacterium (SAR11) and phytoplankton (*Prochlorococcus*). To gain novel insights into microbial interactions that cannot be resolved with bulk-sequencing or culture-based approaches, we utilized flow cytometry with fluorescence-activated single-cell sorting (FACS) and follow-up genome sequencing. FACS separates individual cells based on cell characteristics like the presence of a food vacuole or pigmentation, only co-sorting those microbes that are physically associated and in close proximity with the single targeted eukaryotic cell. This allows selection and sorting of photosynthesising and actively feeding

protist cells directly from field samples and is especially useful to investigate small cells that cannot be hand-picked via microscopy to sequence single cell genomes and analyse the composition of their microbiome.

### **Protists in long-term timeseries**

Microbial interactions can vary depending on seasonality, environmental conditions and could be affected by changes in the water column due to climate change. Long-term timeseries studies are essential for understanding changes in environmental conditions and microbial communities that occur as a consequence of natural and anthropogenic forcing. This enables an investigation of changes in microbial interactions driven by changes in temperature and biogeochemical parameters. The Bermuda Atlantic Time-series Study (BATS) is one of the two oldest ongoing long-term timeseries, started in 1988, located at 31° 40'N, 64°10'W in the Sargasso Sea, roughly 100 km off the shore of Bermuda in the North Atlantic Ocean. Since 1991, the BATS timeseries includes not only basic physical and biogeochemical parameters, but also quantifications primary productivity and carbon export [17], as well as the collection of molecular samples, particularly DNA. Since 2016, the establishment of the BIOS-SCOPE project focused on microbial oceanography and its impact on the carbon cycle has led to an increase in sampling resolution, particularly in depth-resolution. The availability of archived DNA samples harbors the potential to generate a high-resolution dataset of the composition of the microbial community over a 30-year timespan. During this time period, BATS has experienced substantial changes in biogeochemical parameters, including a phase of rapid warming between 2010 and 2020, and a decline in nutrient inventories in surface waters, from increased thermal stratification [17].

### **Thesis Outline**

Here, I seek to explore interactions of eukaryotic phytoplankton and predatory protists with other marine microbes and changes in their occurrence as a consequence of long-term changes in open ocean ecosystems. I employ a variety of single-cell genomics and amplicon-based approaches to gain insights into the type of interactions seen in marine protists and their evolutionary history.

Chapter 2 is centred around bacteria found in physical association with eukaryotic phytoplankton, representing members of the phycosphere and compare its composition to the microbiome of heterotrophic protists. I identify Bacteroidota, Planctomycetota and

Verrucomicrobiota as the three main bacterial lineages found within the phycosphere of various eukaryotic phytoplankton lineages. They are characterized by a vast genomic repertoire of enzymes to break down carbohydrates and encode different systems associated with physical attachment to cells, distinguishing them from smaller, streamlined genomes of symbionts of heterotrophic protists.

In chapter 3, I focus on four uncultured, divergent lineages (two Rickettsiales and two gammaproteobacterial clades) of symbionts of predatory protists, specifically associated with either a marine stramenopile (MAST-3) or a choanoflagellate (*Bicosta minor*). Their evolutionary position enabled insights on the emergence of key metazoan pathogens, including human disease vectors. Sharing common characteristics of symbionts, including small genome sizes and the presence of a type IV secretion system, they show distinct evolutionary trajectories. We observed a Rickettsiales lineage with major genome reduction and key metabolic enzymes indicating an obligate intracellular symbiosis, while a traditional Legionellales lineage displayed extensive acquisition of genes through horizontal gene transfer from distantly related bacterial pathogens. My findings illuminate symbiont evolution in the context of their relationship to hosts, including marine protists as well as animals, including humans themselves.

In chapter 4, I describe findings that suggest that Marinamargulisbacteria, part of the *Candidatus* Margulisbacteria, are likely symbionts of protists. We recovered multiple assemblies of three distinct lineages of Marinamargulisbacteria associated with the choanoflagellate *B.minor*, and two stramenopiles lineages (MAST-4 and MAST-1). Margulisbacteria were previously only known to be associated to multicellular eukaryotes from the Metazoa, with our findings suggesting that their eukaryote-associated lifestyle extends to protists as well, similar to patterns observed in Rickettsiales and Coxiellales. The global distribution of Marinamargulisbacteria across habitats suggests that their host-associated lifestyle is a conserved strategy enabling their success in spreading across ecosystems.

Chapter 5 focused on long-term changes in microbial community composition utilizing a 30-year (Oct. 1991 – Dec. 2021) V1V2 16S rRNA gene amplicon timeseries collected at BATS spanning the surface and mesopelagic open ocean. The combination of V1V2 16S rRNA gene plastid and bacterial amplicons allowed an analysis of co-occurrence between phytoplankton and potentially associated bacteria in the context of a changing environment.

Overarchingly, I observed a decline in relative abundance of phytoplankton in general, including cyanobacteria and clades of eukaryotic phytoplankton, possibly driven by increased stratification leading to more nutrient depleted conditions in the surface layer.

The objectives of this work were to:

1. Characterizing the microbiome of photosynthetic protists on a single-cell basis, comparing it with the microbiome of heterotrophic protists, and predicting the potential interactions between bacteria residing in the phycosphere and their host cell.
2. The diversity of symbionts of heterotrophic protists, their genomic potential, and their evolutionary history in the context of their relationship with pathogens on metazoans.
3. Long term trends in diversity and interactions between microbes in the context of a changing open ocean ecosystem.

This dissertation aims to disentangle inter-domain interactions between bacteria and diverse protist lineages based on culture-independent comparative genomics and timeseries analysis.

## References

1. Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science (1979)* 2015; **347**: 1257594.
2. Azam F, Fenchel T, Field JG, Gray JS, Meyer-Reil LA, Thingstad F. The Ecological Role of Water-Column Microbes in the Sea. *Mar Ecol Prog Ser* 1983; **10**: 257–263.
3. Pomeroy LR. The Ocean’s Food Web, A Changing Paradigm. *Bioscience* 1974; **24**: 499–504.
4. Field CB. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science (1979)* 1998; **281**: 237–240.
5. Stocker R. Marine Microbes See a Sea of Gradients. *Science (1979)* 2012; **338**: 628 LP – 633.
6. Seymour JR, Amin SA, Raina J-B, Stocker R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nat Microbiol* 2017; **2**: 17065.
7. Martin BD, Schwab E. Symbiosis: “Living together” in chaos. *Studies in the History of Biology* 2012; **4**: 7–25.
8. Preer Jr JR, Preer LB, Jurand A. Kappa and other endosymbionts in *Paramecium aurelia*. *Bacteriol Rev* 1974; **38**: 113–163.
9. Fujishima M, Kodama Y. Mechanisms for establishing primary and secondary endosymbiosis in *Paramecium*. *Journal of Eukaryotic Microbiology* 2022; **69**: e12901.
10. Yijing S, C QD, Yuehui T, Siyi Z, Qingyun Y, Zhili H, et al. The Ecology and Evolution of Amoeba-Bacterium Interactions. *Appl Environ Microbiol* 2021; **87**: e01866-20.
11. Maëlle M, Matthias H, Michael W, Marina S, Yousef AK. Amoebae as Training Grounds for Intracellular Bacterial Pathogens. *Appl Environ Microbiol* 2005; **71**: 20–28.
12. Biggs HM. Diagnosis and management of tickborne rickettsial diseases: Rocky Mountain spotted fever and other spotted fever group rickettsioses, ehrlichioses, and anaplasmosis—United States. *MMWR Recommendations and Reports* 2016; **65**.
13. Driscoll TP, I VV, L GM, S LS, A RS, Magda B-S, et al. Wholly Rickettsial! Reconstructed Metabolic Profile of the Quintessential Bacterial Parasite of Eukaryotic Cells. *mBio* 2017; **8**: e00859-17.

14. Zubkov M V, Tarran GA. High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature* 2008; **455**: 224–226.
15. Wilken S, Yung CCM, Poirier C, Massana R, Jimenez V, Worden AZ. Choanoflagellates alongside diverse uncultured predatory protists consume the abundant open-ocean cyanobacterium *Prochlorococcus*. *Proceedings of the National Academy of Sciences* 2023; **120**: e2302388120.
16. Mayers TJ, Bramucci AR, Yakimovich KM, Case RJ. A bacterial pathogen displaying temperature-enhanced virulence of the microalga *Emiliana huxleyi*. *Front Microbiol* 2016; **7**: 892.
17. Lomas MW, Bates NR, Johnson RJ, Steinberg DK, Tanioka T. Adaptive carbon export response to warming in the Sargasso Sea. *Nat Commun* 2022; **13**: 1211.
18. Kwiatkowski L, Torres O, Bopp L, Aumont O, Chamberlain M, Christian JR, et al. Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections. *Biogeosciences* 2020; **17**: 3439–3470.
19. Doney SC, Ruckelshaus M, Emmett Duffy J, Barry JP, Chan F, English CA, et al. Climate Change Impacts on Marine Ecosystems. *Ann Rev Mar Sci* 2011; **4**: 11–37.
20. Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, et al. Determinants of community structure in the global plankton interactome. *Science (1979)* 2015; **348**: 1262073.
21. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. *Science (1979)* 2015; **348**: 1261359.
22. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* 2018; **5**: 170203.
23. Nayfach S, Roux S, Seshadri R, Udway D, Varghese N, Schulz F, et al. A genomic catalog of Earth’s microbiomes. *Nat Biotechnol* 2021; **39**: 499–509.

## Chapter 2

### **Diversity and comparative genomes of prokaryotes associated with heterotrophic and photoautotrophic marine protists**

---

Fabian Wittmers<sup>1,2</sup>, Camille Poirier<sup>2,3</sup>, Lisa Sudek<sup>3</sup>, Rex Malstrom<sup>4</sup>, Alexandra Z. Worden<sup>1</sup>

<sup>1</sup>Marine Biological Laboratory, Woods Hole, MA 02543, USA

<sup>2</sup>Ocean EcoSystems Biology Unit, RD3, GEOMAR Helmholtz Centre for Ocean Research, 24148 Kiel, DE

<sup>3</sup>University of California Santa Barbara, 93106 Santa Barbara, CA, US

<sup>4</sup>DOE Joint Genome Institute, 94720 Berkeley, CA, US



## Abstract

Marine protists are diverse, often uncultured, but occupy crucial positions in the marine food web, including major groups of phytoplankton contributing to carbon fixation, as well as key predatory protists feeding, transporting energy through the food chain, and controlling population sizes. Their relationships with other microbes going beyond classical predator-prey relationships, including their own microbiome, but these more complex interactions remain poorly characterized. We utilized single cell sorting and whole genome sequencing of marine protists to discover bacteria associated with photosynthetic and heterotrophic protists. We observed a clear separation in taxonomy and genome characteristics for symbionts of protists with different lifestyles. Putative symbionts from the Alphaproteobacteria and Gammaproteobacteria associated with heterotrophic protists included four new lineages with small genome sizes related to either the Rickettsiales or the Coxiellales, orders which include many intracellular symbionts and pathogens of metazoans. The microbiomes associated with photosynthetic protists were dominated by members of the Planctomycetota, Verrucomicrobiota and Bacteroidota. These bacteria had larger genomes, encoding a more diverse repertoire of non-essential metabolic functions, especially centred around carbohydrate-active enzymes (CAZys). Most of these members of the so-called phycosphere also encoded either secretion systems, pili or flagella. Such structures enable varying degrees of motility and attachment to surfaces, suggesting an extracellular lifestyle involving periods of attachment to the eukaryote. Families of CAZys specific to break down modified glycoproteins on the phytoplankton's surface, indicating that these lineages might not only utilise carbohydrate secreted by phytoplankton during photosynthesis, but could instead attack the cell-wall itself, potentially depending on the host they attach to and variability in environmental conditions. The diversity and evolutionary position of the novel symbiont-host relationships of marine protists uncovered could enable a better understanding of the diverse cross-domain interactions and flow of resources occurring in the microbiome of marine protists generally, including the phycosphere of marine phytoplankton and the microbiome of heterotrophic protists alike.

## Introduction

Marine phytoplankton are a cornerstone of the global carbon cycle, contributing roughly 50% of the annual global carbon fixation [1]. During photosynthesis, phytoplankton produce excess dissolved organic matter (DOM), part of which ends up in its surrounding environment, through a variety of active or passive processes. The amount of DOM released by cells can vary greatly (getting as low as 2%) but can reach up to 50% of primary production in diatom cells [2]. Based on its chemical composition, phytoplankton releases a wide variety of organic compounds, mostly comprising carbohydrates, lipids, nitrogenous compounds, and organic acids [3]. The passive or active export of DOM causes its accumulation in the microscale environment surrounding the phytoplankton cell, which is referred to as the phycosphere [4]. It contains components that likely serve as chemical attractants for bacteria [5] and this DOM-rich microenvironment provides desirable conditions for heterotrophic bacteria which can utilize DOM as an energy source, e.g., through Carbohydrate-Active EnZymes (CAZys) that enable the breakdown and subsequent respiration of carbohydrates or glycosylated proteins [6]. The presence of these heterotrophic bacteria, essentially the phytoplankton cell's microbiome, in turn benefits the phytoplankton cell by complementing their metabolic needs [7]. This leads to a symbiotic relationship between phytoplankton and the bacterial members inhabiting its phycosphere, benefiting both sides. These same types of relationships could be hypothesized for heterotrophic protists.

In theory, it is thought that phytoplankton, at least partially, regulate the amount and composition of DOM exudation in response to varying nutrient conditions and limitations [2], potentially to attract bacteria that benefit them during these conditions. DOM release increases when phytoplankton cells are nutrient limited [8–10] which is generally explained by changes in metabolism as a response to nutrient stress leading to increase in metabolic waste (i.e., DOM, that is subsequently released) – the overflow hypothesis, but an alternative explanation could be increased DOM secretion to attract beneficial bacteria that might benefit the phytoplankton cell during nutrient-limiting conditions [7]. For example, members of the *Rhodobacterales* (Alphaproteobacteria) are frequently found to respond to phytoplankton blooms and are common in phytoplankton cultures [11] and can cross-exchange vitamins with phytoplankton, compensating their inability to synthesise vitamins like vitamin B12 themselves [12]. They have therefore been proposed as members of the phycosphere and their genomes encode systems that could potentially enable attachment to their 'host', but this physical co-association has not been shown [13].

The study of the interaction of phytoplankton and bacteria has been centred around diatom culture-systems, because of fast-growing model organisms and their global importance for primary production in the ocean [7]. Conceptually, most of these studies focus on the release of compounds by phytoplankton and their uptake by bacteria in co-culture [2, 7]. This follows the current approach to measure metabolomics (i.e., DOM composition) in the ocean, focusing on deciphering the diversity of metabolites present, challenged by the low standing stock concentrations and hard to measure, often fast turnover rates [14]. Essentially, the interaction between phytoplankton and their microbiome is studied as a free market exchange, focusing on production and consumption of compounds, but not looking into the dynamics of co-association within the phycosphere. Our understanding of this micro-environment is still very confined, mainly limited by the complexity of any, especially in-situ, measurements at this scale and the lack of insights into physical associations.

Comparably, the study of the microbiome of heterotrophic protists, especially those lineages that dominate the open ocean, playing a major role in the marine food-web, is still in its infancy, with even fewer insights than those gained for the phycosphere. Our understanding of the microbiome of heterotrophic protists is mostly confined to those groups that have large cell sizes, and cultured representatives, for example Rhizaria and some Dinoflagellates, with most of those studies focusing on kleptoplasty or symbioses with photosynthetic entities [15–17]. Virtually nothing is known about microbiomes and interactions with nanoflagellates (2-20  $\mu\text{m}$ ) that are important to grazing in the open ocean [18, 19]. Bulk metagenomics, generating **metagenomics assembled genomes** (MAGs) and amplicon surveys have revealed various bacterial lineages that correlate with phytoplankton and/or heterotrophic protists and have highlighted the importance of biotic interactions, e.g., in overcoming nutrient limitation, but these studies cannot gain insights into physical association [20, 21]. We therefore have very limited insights into true interactions in the wild between photosynthetic or heterotrophic protists and bacteria alike, requiring the use of alternative approaches to gain insights into this gap.

We utilized flow cytometry with fluorescence-activated single-cell sorting (FACS) with follow-up whole genome sequencing to gain new insights into some of the many open questions that remain for the attached microbiome of marine protists. Flow cytometry enables distinction of marine entities based on size and pigmentation and was prominently utilized in oceanography in the discovery of the abundant cyanobacterium *Prochlorococcus* [22]. Since it separates individual cells, it can be used to sort these cells individually based on pigmentation or staining of food vacuoles and allows selection of photosynthesising and

actively feeding protist cells from field samples. This is especially useful to investigate small cells that cannot be hand-picked via microscopy to sequence single cell genomes to analyse the composition of their microbiome.

## Material & Methods

### Single cell sorting and amplification

Sorting was performed using freshly collected marine photic zone samples that had been passed through a 30  $\mu\text{m}$  nylon mesh upon collection and subsequently concentrated by gravity over a 0.8  $\mu\text{m}$  Supor filter (Pall Corporation). Single cells were sorted from environmental samples collected from the North Pacific (Monterey Bay, 733 cells) and a mangrove ecosystem (Curaçao, 48 cells). Samples were collected from the upper photic zone between the surface and 42m during multiple sampling campaigns between 2014 and 2019. An InFlux high speed sorter (BD) equipped with 488 nm and 355 nm lasers and a 70  $\mu\text{m}$  nozzle, operating on the high purity Single-Cell sorting mode, with sterile nuclease-free 1x PBS as sheath fluid (ThermoFisher cat# AM9625 and AM9932), was used in all cases. First, samples were stained for 10-20 min in the dark with LysoTracker Green DND-26 (25 nM, final concentration) and/or LysoSensor Blue DND-167 (1  $\mu\text{M}$ , final concentration) or no staining. Unstained controls were run for all samples such that positive signal from the stain was always discerned as compared to the control. Forward Angle Light Scatter (FALS), and, depending on the protist population of interest, red fluorescence from chlorophyll-a (692/40 nm bandpass), green fluorescence from LysoTracker (520/35 nm bandpass), and blue fluorescence from LysoSensor (435/30 nm bandpass) were used as selection criteria, while all other particles were selected for exclusion. Based on their sorting gate, sorts were operationally classified into those targeting putatively heterotrophic protists (758 cells) and photoautotrophic protists (23 cells).

Immediately prior to sorting plates were illuminated with UV irradiation for 2 min. Negative and positive controls (for subsequent genome amplification reactions) were performed on all plates by leaving a subset of wells empty, and by having a column of wells receiving multiple cells within each 384 well plate, respectively. Additionally, the drop delay and sort quality were regularly controlled by sorting a known number of fluorescent bead standards and counting them by epifluorescence microscopy. After sorting, plates were covered with foil seals and frozen at  $-80^{\circ}\text{C}$ . Single-protist or multi-protist cell sorts were subjected to Multiple Displacement Amplification using the repliG or WGA-X enzyme, as previously described (Needham et al, 2019). Sequencing was performed on the Illumina platform, resulting in an average of 2,617,247 raw reads per sequenced well.

### Single-protist-well DNA sequence assemblies and ipaMAG recovery

Demultiplexed paired-end reads were quality controlled and trimmed using trimmomatics (v0.39; [17]), with a trimming cutoff quality score  $<3$  or when a 25 bp sliding window average was  $<30$ . Assembly was performed using Spades (v3.15.5; [18]) with the single cell option ``-sc`` and kmers 21, 33, 55, 77, 99 and 127, considering both paired and unpaired reads that passed quality-control. Samples that could not be assembled through spades single cell mode using a maximum of 180 Gb RAM were instead assembled using megahit (v1.2.9; [19]) with the `—presets meta-sensitive` settings, considering both paired and unpaired trimmed reads that passed quality-control. All assembled contigs were filtered and only those  $> 1$  kb were retained. Sequence alignment maps for filtered contigs were generated using bowtie2 (v2.4.1; [23]), with default setting and unaligned reads removed from the alignment map. Sequence alignment files were sorted and compressed using samtools (v1.6; [24]) `‘view -bS’`, `‘sort’` and `‘index’`. Complete or fragments of 16S or 18S rRNA genes were identified using checkM’s (v1.2.2; [25]) `‘ssu_finder’` subcommand with default settings.

Prokaryotic genomes were recovered using Anvi’o (v8; [26]), with initial clustering of contigs using `‘anvi-profile’` and  $>95\%$  nucleotide identity. Cellular single copy marker genes for interactive genome completeness estimate were identified using `‘anvi-run-hmms’` and the default database. Preliminary taxonomy of the recovered genes was predicted at the protein level using `‘anvi-run-scg-taxonomy’` with default settings. Additionally, `‘anvi-run-hmms’` was run using a set of 47 HMMs for marker genes of eukaryotic giant viruses (NCVOGs; [27]). Profiles were manually inspected for anomalies in GC% content and tetranucleotide frequency and inspected interactively, binning potential prokaryotic genomes. All bins of GC anomalies were initially screened for phage/viral contamination using the following approach: Proteins were queried against NCBI nr using diamond blastp (v2.14; [28]) with E-value  $< 1 \times 10^{-4}$  and output format 6, including an additional column for subject NCBI taxonomy ID. IDs were converted into taxonomy strings using NCBI taxid2lineage information (obtained through: [github.com/zyxue/ncbitax2lin](https://github.com/zyxue/ncbitax2lin)). Up to 100 query hits per protein were considered and hits were summed up for each contig individually when calculating the percent of viral hits. Contigs were removed from the raw bin when at least 2% of hits were classified as viral. Virus-clean bins were assessed using checkM’s `‘lineage_wf’` subcommand with default settings and all bins identified as genomes of 10% completeness or higher were hereafter referred to as ipaMAGs (**i**ndividual **p**rotist **a**ssociated **M**etagenome **A**ssembled **G**enome). Genomes were not called single-cell amplified genomes (SAGs) as these assemblies could be based on multiple prokaryotic genomes associated with the same

eukaryotic cell sorted and sequenced. Taxonomy of ipaMAGs was predicted using GTDB-Tk's (v1.0.2; [29]) 'identify', 'align' and 'classify' consecutively, with GTDB version 214.1 as the reference database and default settings when meeting GTDB-Tk minimum classification cutoffs.

Proteins were predicted using 'prodigal' (v2.6.3; [30]) for each ipaMAG of medium or high quality (>50% completeness, <10% contamination; following definitions as in [31]) individually using default settings. Pairwise average amino-acid identity between all ipaMAGs was calculated using the aai.rb script of the 'ecogenomics' collection (downloaded 20 June 2023; [32]) with default settings. Proteins for all ipaMAGs were annotated using emapper.py (v2.0.1; [33]) against the eggNOG database (v5.0; [34]) with default settings. Additionally, Pfam domains (v35; [35]) were identified using hmmscan (v3.3.2; [36]) with an E-value of  $< 1 \times 10^{-10}$ . Homologies against NCBI nr database (downloaded 28 Mar 2023) were identified using Diamond blastp (v2.14; [28]), requiring a minimum query coverage of 50 bp and an E-value of  $< 1 \times 10^{-04}$ . For identification of secretion systems, MacSyFinder (v2.1; [37]) was run with default settings using both the 'TXSS' and 'TFF-SF' models (downloaded 28 Mar 2023). CAZys (Carbohydrate-Active enZYmes) were annotated using hmmscan vs an HMM model version of the CAZy database (Carbohydrate Active Enzymes database; [38]) and filtered using the CAZydb hmmscan-parser script (both available through dbCAN3: <https://bcb.unl.edu/dbCAN/>; CAZydb v.11). High quality ipaMAGs (>90% completeness, <10% contamination) were additionally annotated against the KEGG database using the GhostKOALA webserver ([kegg.jp/ghostkoala/](http://kegg.jp/ghostkoala/); [39]) with default settings.

### 16S rRNA gene screening & phylogenetic placement

For subsequent phylogenetic placement, a global 16S rRNA gene reference phylogeny was constructed using the SILVA database (v138.1 SSURef NR99). The nucleotide alignment was randomly subsampled to 10% of the original number of taxa in the alignment (51,423 gene sequences). The alignment was then trimmed using trimAl (vX; [40]) with '-gt 03', and redundant sequences were reduced to a single representative using 'seqkit rmdup', resulting in the removal of 2617 gene sequences. A phylogenetic tree was then reconstructed from the resulting alignment using FastTree (v2.1.8; [41]), initially masking for non-homologous regions and using the GTR model and '-fastest'.

All assembled contigs were screened for 16S or 18S rRNA gene-containing sections using phyloflash (v3.4.2; [42]) with default settings. Sequences identified as containing a 16S

rRNA gene fragment were trimmed to extract the respective rRNA section and blasted against SILVA (v138.1 SSURef NR99) using blastn. Sequences with a top hit to either a mitochondrial, chloroplast or eukaryotic sequence (E-value  $< 1 \times 10^{-04}$ ) were removed. 624 sequences (min 50 bp) were subsequently aligned against the modified SILVA reference alignment using SINA (v1.7.2; [43]) with default settings. Sequences were then placed on the previously constructed global 16S rRNA gene phylogeny using epa-ng (v0.3.8; [44]) with default settings. Each sequence's last common ancestor-placement was calculated using gappa edit accumulate (v0.8.4 [45]) with a minimum likelihood weight ratio threshold of 0.9. The phylogeny and placements were visualized in R using ggtree (v3.6.1; [46]).

#### Detection of symbiont-lineages in global ecosystems

16S rRNA gene sequences recovered from sort assemblies were used as queries against V1-V2 16S rRNA amplicon sequence variants (ASVs) in published datasets from the Bermuda Atlantic Timeseries Study (BATS) in the North Atlantic [47], the Caribbean Sea [48], the Mediterranean Sea [49] and the Indian Ocean [50, 51]; and eastern North Pacific [52, 53] were combined and expanded with additional amplicon sequencing samples. For Monterey Bay, a new feature table and representative sequences, combining previously published and new samples, were generated as previously described [47]. Classification as identical to the query required a 100% nucleotide match for the complete V1-V2 region amplified and classified as the same lineage required  $\geq 97\%$  nucleotide identify, and additionally requiring a minimum 90% alignment of query and subject.

#### Functional profiling of recovered bacterial genomes and reference genomes

For comparison of functional capacity of high-quality genomes recovered here with reference genomes, all genomes were annotated as follows: Proteins were predicted and annotated using GhostKOALA and HMMSCAN against Pfam and the CAZy hmm db respectively, as described above. Protein counts for CAZys were accumulated for the five major functional families discriminated in the database. Proteins were annotated as ATP/ADP translocases based on the Pfam domain PF03219 with an E-value of  $< 1 \times 10^{-25}$ . Additionally, tRNAs were identified using ARAGORN (v1.2.41; [54]) with default settings. Proteins were assigned to lipid metabolism (beta-oxidation, lipid biosynthesis and lipopolysaccharide synthesis), nucleotide metabolism (purine biosynthesis, pyrimidine biosynthesis), respiration (glycolysis, oxidative phosphorylation, TCA cycle), amino acid metabolism and vitamin

synthesis based on their assigned to metabolic modules within KEGG. Rhodopsin and retinal biosynthesis protein annotations are based on Pfam domains for rhodopsin (PF01036.18), phytoene synthase (PF00494.19), GGPP synthase (PF00348.17), beta-carotene 15,15' dioxygenase (PF15461.6) and lycopene cyclase (PF05834.12) with a cutoff of  $< 1 \times 10^{-10}$ . The number of proteins for each functional category (see columns in Fig.3A) was z-score normalized for visualization purposes.

#### Data and code availability

Raw amplicon sequencing reads for the Monterey Bay environmental samples are available in NCBI under BioProject PRJNA1107937. Representative ASVs and the raw feature table are deposited in FigShare ([doi.org/10.6084/m9.figshare.25751283.v1](https://doi.org/10.6084/m9.figshare.25751283.v1)). Raw reads for single-cell sorts and the highest quality ipaMAG of each lineage are available via NCBI BioProject PRJNA1061664.

## Results & Discussion

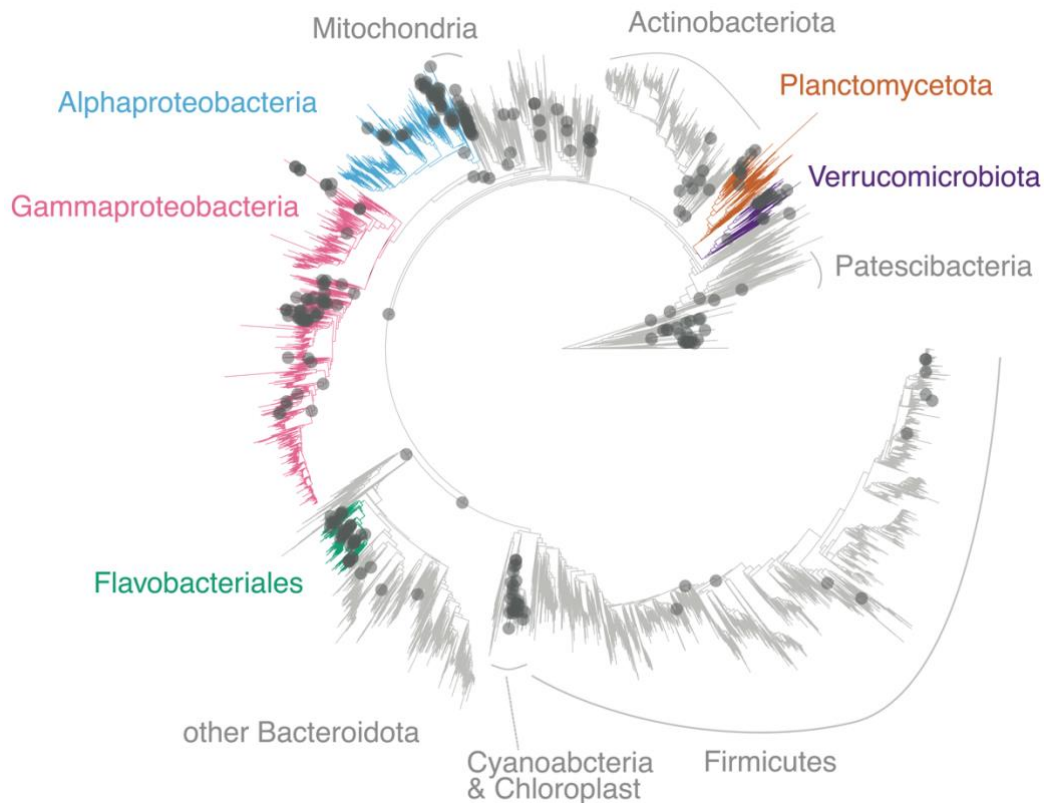
To separate eukaryotic and prokaryotic assemblies found within sequenced single-cell sorts, we initially processed all sequences from each sample together. 781 eukaryotic sorts produced assemblies larger than 250 Kb, with an average assembly size of  $10.55 \pm 10.90$  Mb and a cumulative size of 8235.883 Gb. SAGs were sorted from environmental samples collected from the North Pacific (Monterey Bay, 733 cells) and a mangrove ecosystem (Curaçao, 48 cells). All samples for cell sorting were collected from the upper photic zone between the surface and 42 m during sampling campaigns between 2014 and 2019. Based on their sorting gate, cells have been operationally classified into those targeting putatively heterotrophic protists (758 cells) and photoautotrophic protists (23 cells).

The dataset also comprises 77 sorts for which multiple cells were directed to the same tube or well, hereafter referred to as population sorts. These had 15-1000 cells each that were then sequenced as a whole, resulting in “targeted” metagenomes. The latter had an average assembly size of  $141 \pm 122.56$  Mb. Based on (partial or complete) 18S rRNA gene sequences recovered from 510 SAGs (65%), the dataset comprises eukaryotes from seven supergroups [55]. The location and / or timing of the sampling rendered many choanoflagellate sorts ( $n=327$ , where  $n$  represents the number of sorts with 18S rRNA gene fragments  $>300$  bp) and these were almost entirely represented by a single species (*B. minor*). Other well-represented lineages include dinoflagellates ( $n=61$ ), stramenopiles ( $n=39$ ), Telonemia ( $n=26$ ), Rhizaria ( $n=17$ ) as well as Picozoa ( $n=14$ ).

### Marine protists associated bacteria are dominated by five lineages

Analysis of the 16S rRNA genes was performed in order to access the taxonomic composition of the prokaryotic DNA sequenced from each sort. This provided mitochondrial and chloroplast (phytoplankton sorts only) sequences of eukaryotic organelles for many of the sorts, as expected. It also revealed diverse bacterial sequences. Phylogenetic placement of fragmented and complete 16S rRNA genes from non-organellar sources, indicated that the detected bacteria primarily consisted of members of the Alpha- and Gammaproteobacteria, Planctomycetota, Bacteroidota (especially Flavobacteriaceae) and Verrucomicrobiota (Figure 1). Whole genome sequencing allowed us to potentially bin genomic bacterial DNA that assembled, allowing further insights into the genomics underlying the physical associations we detected. A portion of eukaryotic single cell sort assemblies were comprised of multiple clusters of contigs with distinct GC content far exceeding typical variation observed within a

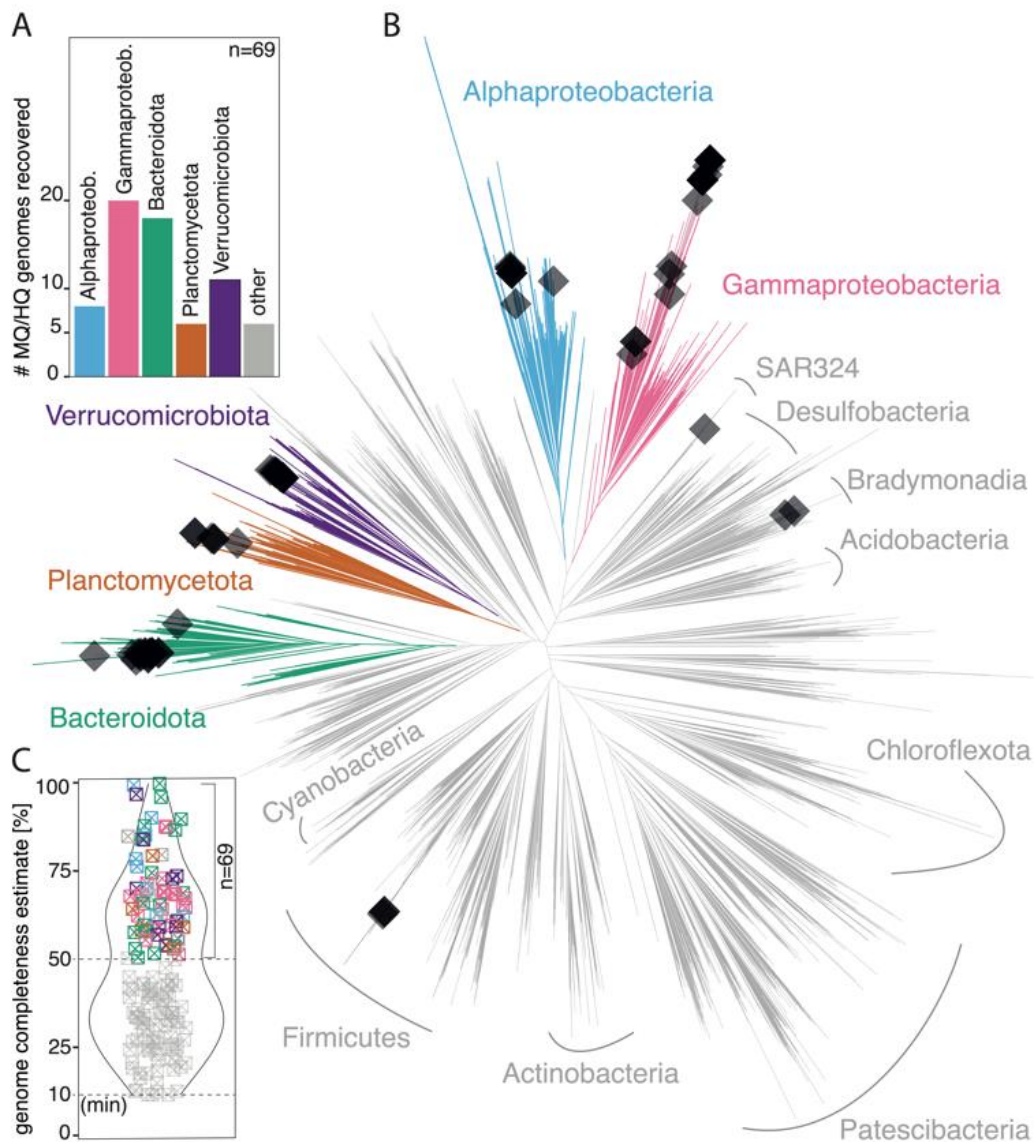
single genome. Although genomic and organelle-derived DNA differs in GC-content, some of these distinct GC regions far exceeded the genome size of organelles and instead putatively comprised distinct prokaryotic DNA. We separated bacterial DNA from host DNA based on GC-content, assembly quality (with prokaryotic DNA assembling better than eukaryotic DNA in mixed samples) and the presence of bacterial single-copy marker genes.



**Figure 1: Bacterial lineages found in co-association with marine protists.** Maximum likelihood reconstruction of SILVA138.1-based 16S rRNA gene phylogeny under the GTR model with 2113 positions. Each circle indicates the placement of a 16S rRNA gene fragment (min 50 bp) originating from a eukaryotic cell-sort placed on the node with a minimum likelihood-weight ratio of 0.9.

We manually inspected clustering of assembled sequences in all sorts separately, required because automated genome recovery tools require differential coverage, which is skewed in single-cell genomics by the usage of MDA, relying on an enzyme that amplifies DNA before sequencing. This led to the identification of 232 putative prokaryotic DNA clusters that we separated from the eukaryotic DNA present in the assembly. One-hundred-fifty-eight of these, recovered from 100 distinct sorts, were of sufficient quality to estimate genome completeness and contamination. Out of a total these 158 ipaMAGs, 69 were categorized as prokaryotic assemblies of medium or high quality (according to [31]; min. 50%

completeness, max. 10% contamination), and kept for downstream analysis (Figure 2). These ipaMAGs represented 13 distinct bacterial orders and averaged  $2.13 \pm 1.31$  Mb in genome size. The majority (90%) were classified as either Alpha- or Gammaproteobacteria, Bacteroidota, Planctomycetota, or Verrucomicrobiota (Figure 2A).



**Figure 2: Uncultivated bacteria living in physical co-association with wild protists and presence across tropical and subtropical oceans.** (A) Evolutionary positions of bacterial ipaMAGs. Placement is based on ML analysis of a concatenated alignment of 120 proteins encoded by single copy universal genes with 5036 positions in the final masked alignment and any recovered ipaMAG with estimated completion >50% and contamination <10% was included. Branches are labelled (with corresponding color coding) based on GTDB taxonomy. (B) the total number of ipaMAGs recovered for each lineage in association with a protist. (C) Phylum level taxonomic distributions of all bacterial sequences as a function of genome completion (only those >10% complete are included).

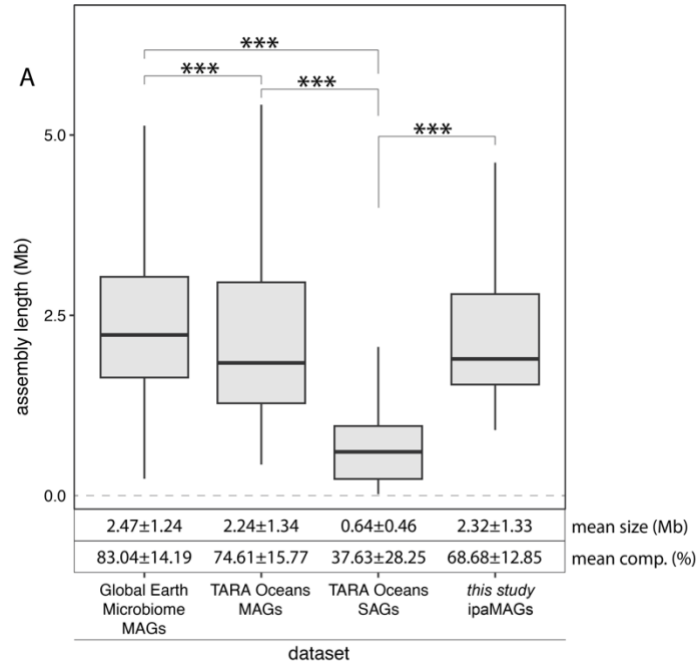
We next used genome-based taxonomy prediction, using a set of 120 marker genes that is conserved throughout all bacteria with a single copy to predict taxonomy of all ipaMAGs recovered, independent of the presence of a 16S rRNA gene. The resulting taxonomy aligned with the main lineages observed in 16S rRNA gene-based phylogenetic placement of the protist-associated bacterial lineages (Figure 2). This suggests that our method of genome assembly and manual ipaMAG recovery is not biased against certain lineages, since ipaMAGs from all major clades present in the 16S rRNA gene placement were represented by genomes herein as well. Across all recovered ipaMAGs, high quality genomes were assembled for ten lineages, representing three Bacteroidota, two Verrucomicrobiota, one Planctomycetota and two Gamma- and Alphaproteobacteria clades, respectively (Table 1). On average, these HQ ipaMAGs were  $2.0 \pm 1.1$  Mb in size, comparable to the size of an average MAG from publicly available (marine and cross-ecosystem) datasets. At the same time this exceeded the average size and quality of previous (marine) studies recovering prokaryotic genomes from SAGs, mainly due to lower average genome completeness in the latter (Figure 3). This could be explained by the presence of multiple clonal cells of the prokaryotes recovered being associated with the single sorted eukaryotic cell. This would increase the availability of identical genomic DNA fragments for sequencing, greatly increasing assembly quality compared to the single prokaryotic cells sorted during previous efforts [56]. Additionally, those efforts didn't comprise manual curation of assemblies or removal of viral sequences, increasing possible contamination. The presence of (almost) complete ipaMAGs herein also implies that these lineages are unlikely to represent prey organisms of the associated heterotrophic protists as one would rapidly degrading genomes during the prey-breakdown process. Their taxonomy further suggests that they are not among the more abundant known prey of predatory protists [18].

| genome ID      | clade                 | group            | compl. | cont. | size_Mb | est. G. size | # contigs | # prot. | % GC  |
|----------------|-----------------------|------------------|--------|-------|---------|--------------|-----------|---------|-------|
| ipa-MAG-0934   | Simplirickettsiaceae* | Alphaproteobact. | 100    | 0.0   | 0.97    | 0.97         | 15        | 937     | 36.62 |
| ipa-MAG-0929   | Gamibacteraceae       | Alphaproteobact. | 91.5   | 0.0   | 1.69    | 1.85         | 168       | 1651    | 36.76 |
| ipa-MAG-0156   | Marinicoxiellaceae*   | Gammaproteobact. | 87.3   | 0.6   | 1.70    | 1.93         | 35        | 1484    | 44.74 |
| ipa-MAG-0636   | Protistiellales*      | Gammaproteobact. | 87.1   | 0.3   | 1.40    | 1.60         | 87        | 1341    | 40.90 |
| ipa-MAG-0027-7 | UBA4466               | Bacteroidota     | 99.5   | 8.0   | 3.30    | 3.07         | 352       | 3260    | 39.04 |
| ipa-MAG-0015   | Patiriosocius         | Bacteroidota     | 89.3   | 4.2   | 2.07    | 2.22         | 102       | 1990    | 35.48 |
| ipa-MAG-0027-1 | CAJQQY01              | Bacteroidota     | 86.3   | 3.7   | 2.63    | 2.92         | 234       | 3674    | 32.89 |
| ipa-MAG-0691   | SW10                  | Verrucomicrobia  | 83.5   | 1.4   | 3.54    | 4.17         | 554       | 3247    | 49.49 |
| ipa-MAG-0028-5 | Rubritalea            | Verrucomicrobia  | 96.4   | 2.0   | 3.37    | 3.43         | 217       | 3150    | 49.46 |
| ipa-MAG-0028-7 | Mariniblastus         | Planctomycetes   | 78.9   | 1.4   | 4.27    | 5.32         | 352       | 3408    | 48.40 |

**Table 1: Assembly statistics for ipaMAGs from highest quality lineages recovered.** Clades marked (\*) represent uncharacterized lineages that have been named herein, see chapter 3.

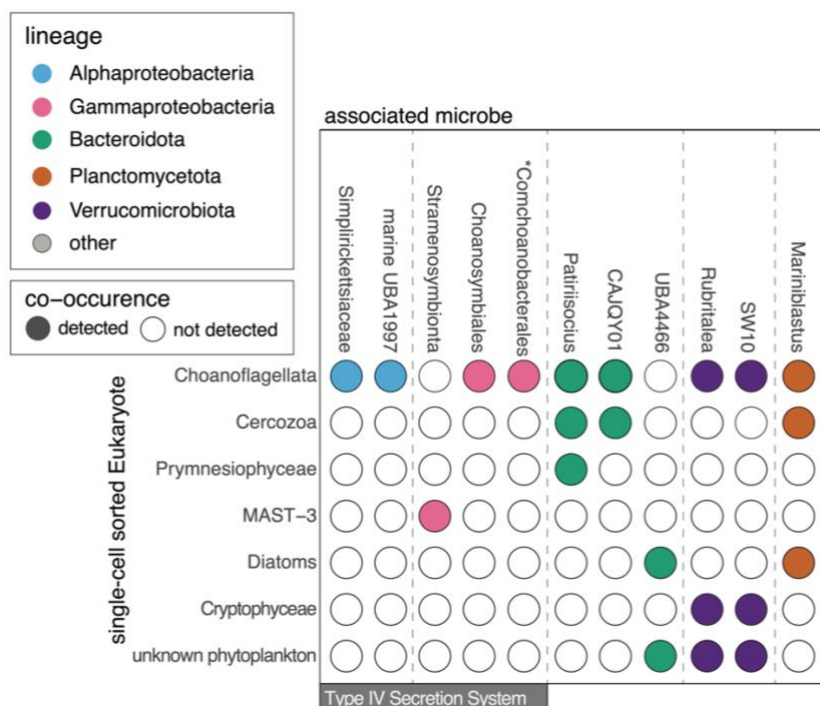
### Variations in host range, genome size and GC content

The representative ipaMAGs showed different patterns with respect to the hosts they were physically associated with (Figure 4). However, while capturing much new information on uncultivated bacteria and their physical association with different protists, the design of sampling itself did not allow for conclusions of the type of interactions. For this reason, deeper investigation of genomic characteristics was employed to reveal evolutionary and functional aspects of the different lineages. This type of information is valuable for understanding evolutionary trajectories, potential host interactions, or functional traits and insights into their lifestyle (see below). Lineages with smaller genomes can evolve through genome streamlining. In free-living cells, it can take place in a larger context of reduction in cell size and complexity as a Darwinian strategy for success in nutrient-limited environments [57]. In symbionts, especially those with intracellular lifestyles, genome size reduces as a consequence of random genetic drift in which the isolation of the symbiont in its host leads to the loss of functionality by chance which can be compensated, e.g., by obtaining essential metabolites, or energy in the form of ATP from their host instead [52]. The stable host environment essentially provides the opportunity to ‘lose’ functions that are redundant with their host organism without facing selective pressures against this loss, i.e., relaxed selection [58]. Smaller genome size is often in line with a reduction in GC-content, driven by less non-coding, GC-rich regions in these genomes [59], energetic efficiency of low-GC replication and maintenance [57, 60], accumulation of AT-mutations as a consequence of higher mutation rates observed in intracellular symbionts [61], and a series of other factors shaping smaller genomes through similar evolutionary drivers.



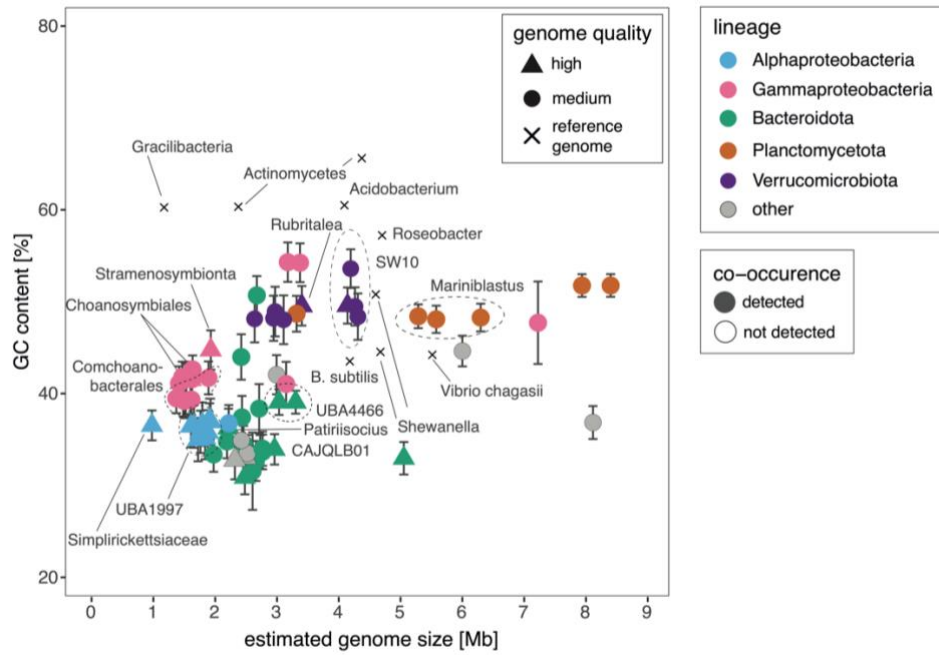
**Figure 3: Comparison of ipaMAG size with previously published data.** (A) Comparison of average size of MAG or SAG assembly size between the Global Earth Microbiome (global cross-ecosystem synteny dataset) [62], TARA Oceans MAGs [63] and TARA Oceans SAGs [56] and ipaMAGs assembled herein. Asterisks indicate p-values obtained through pairwise t-tests with Bonferroni correction (\*\*\*) =  $p < 0.0001$ .

Estimated genome size of ipaMAGs, their GC-content, and broad scale taxonomy was consistent with general trends reported for cultivated reference symbionts and free-living bacteria that have been genome sequenced [57, 59] (Figure 5). Specifically, the alphaproteobacterial and gammaproteobacterial ipaMAGs had significantly ( $p < 0.001$ ) smaller average estimated genome sizes ( $1.7 \pm 0.4$  Mb and  $2.1 \pm 1.4$  Mb, respectively) than the Verrucomicrobiota ( $3.5 \pm 0.7$  Mb) and Planctomycetota ( $6.2 \pm 1.9$  Mb) recovered herein. Similarly, the GC-content of Alphaproteobacteria ( $36.1 \pm 0.7\%$ ) and Gammaproteobacteria ( $42.7 \pm 4.5\%$ ) was lower than that of Verrucomicrobiota ( $49.3 \pm 1.6\%$ ) or Planctomycetota ( $49.5 \pm 1.8\%$ ). Not strictly following the correlation between genome size and GC-content, the Bacteroidota had an average genome size of  $2.8 \pm 0.7$  Mb while having an average GC-content similar to that of Alphaproteobacteria, although with more variability ( $36.2 \pm 4.9\%$ ). Genome streamlining as a consequence of an intracellular lifestyle is not the only driver of GC-content in bacteria, with other selective pressures like resource competition, acting for example on Bacteroidota as members of the phycosphere, affecting the same genomic signatures, explaining the range of genome size for low-GC% lineages observed in microbial communities [57].



**Figure 4: Eukaryotic host association of prokaryotic lineages with high quality ipaMAGs.** Co-association between eukaryotic and prokaryotic organisms in sequenced single cell sorts. Coloured circles indicate that both the eukaryote lineage as well as the associated prokaryotic microbe were detected in the same single cell sort, based on the detection of either an ipaMAG or (partial or complete) 16S rRNA gene sequence.

For ipaMAGs with smaller genomes, we observed a decrease in the variability of the eukaryotic lineage the prokaryotes were associated with. Those with the smallest genomes were identified as possibly representing obligate and/or intracellular symbionts that have specialized for a distinct host lineage (Figure 4). Four symbiont lineages identified within this study with high-quality ipaMAGs phylogenetically placed within Alpha- or Gammaproteobacteria showed specificity in their association with only one protist, indicating a more specialized lifestyle than the Planctomycetota, Bacteroidota and Verrucomicrobiota detected herein. This was additionally supported by the detection of complete type IV secretion systems within all Alpha- and Gammaproteobacteria lineages of high quality (Figure 4), indicative of a host-associated, potentially intracellular, and pathogenic lifestyle (See chapter 3).



**Figure 5: Relationship between GC content, genome size and taxonomy.** Estimated ipaMAG size as a function of average GC content. Error bars indicate GC-variability over a sliding window of 5 Kb. Genomes labelled with 'x' represent reference genomes not recovered in this study but added for context.

### General functional trends

Genome annotation enables the characterisation of functional traits, such as motility and chemotaxis, pathways for degradation or production of specific compounds, or systems associated with secretion of toxins or other proteins in pathogenic symbiont-host associations. The functional analyses allow development and testing of specific hypotheses regarding the lifestyles of the taxa represented by the ipaMAGs. Physically co-association might arise from intracellular symbiosis, either mutualistic or pathogenic, or extracellular attachment to the eukaryote. Extracellular interactions could represent parasitism or even predation, where the bacterium feeds on the eukaryote, or a mutualistic interaction benefiting both. More specialized interactors will have mechanisms for exchanging effector proteins with their specific host, while more generalistic bacteria will encode a more diverse metabolic repertoire to adapt to changes in availability of resources. Subsequently, bacteria associated to the phycosphere of various eukaryotic phytoplankton lineages relying on the increased availability of DOM in that microenvironment should encode more CAZymes (enzymes to modify and break down carbohydrates, a major component of DOM), than pathogens putatively living within a eukaryotic host. If cells not physically attached are not co-sorted, as previously hypothesized [7], then extracellularly associated bacteria must encode structures

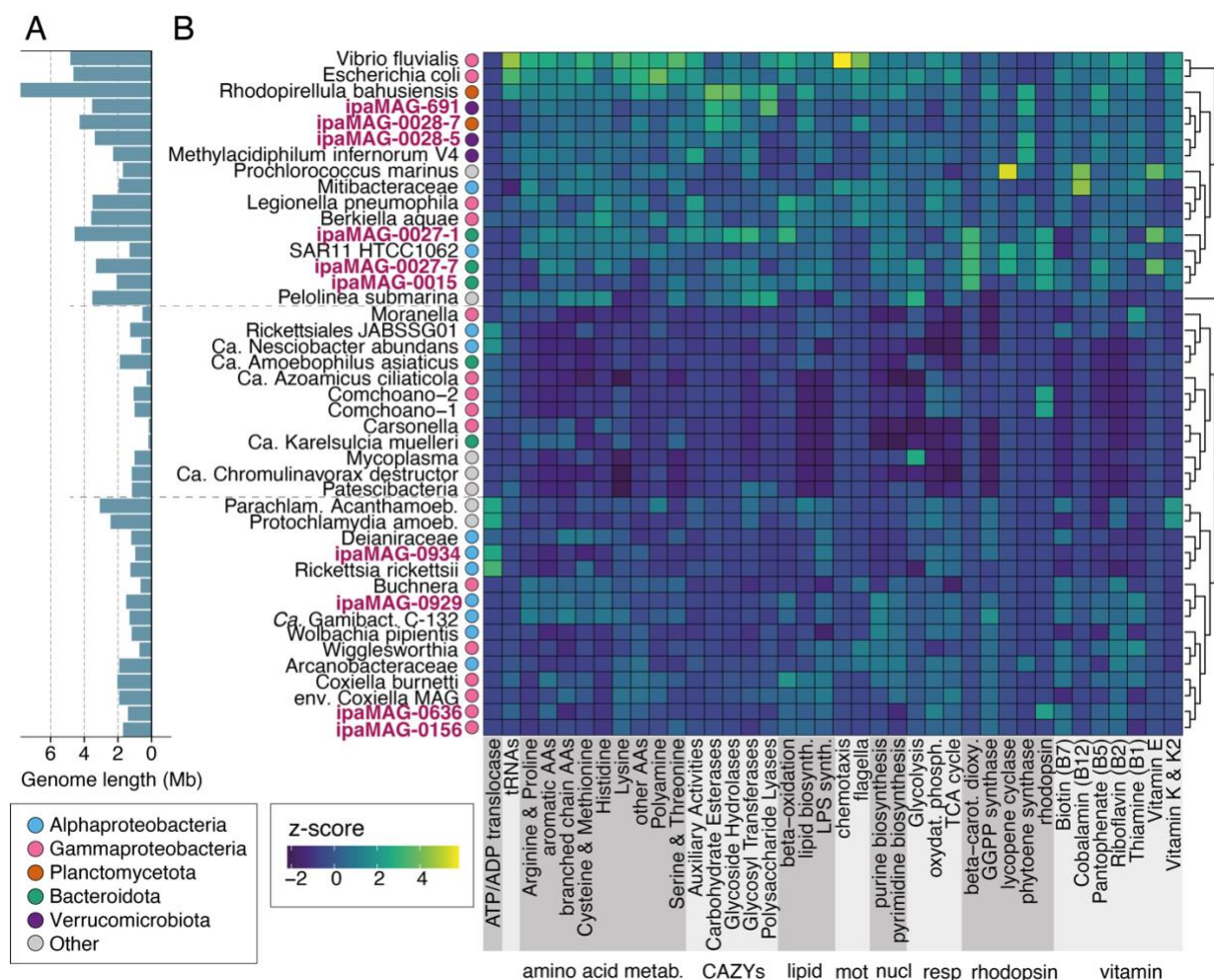
enabling to attach to the surface of the respective eukaryote. Collectively, this type of knowledge gained through functional annotations can then be applied to define the lifestyle or type of relationship they were in with the associated eukaryote.

ipaMAGs with a larger genomic repertoire and free-living capacity, including all Planctomycetes, Bacteroidota and Verrucomicrobiota lineages we recovered, showed the ability to associate with protists of diverse lifestyles and representing lineages placed within different eukaryotic supergroups. The prokaryotic community associated with eukaryotes emitting a photosynthesis-pigment signal were generally more complex than the ipaMAGs associated with putative heterotrophic protists. As opposed to heterotrophy-targeted eukaryotic SAGs, where recovery of more than one associated prokaryote genome was uncommon (15.5% of sorts with at least 1 prokaryotic bin), 3.81 ipaMAGs were found associated with photosynthetic protist SAGs on average.

#### Bacteroidota, Verrucomicrobiota and Planctomycetota as members of the phycosphere

Bacteroidota, especially members of the Flavobacteriaceae family are commonly found to interact with diatoms on a global scale, based on co-occurrence analysis [20], and this interaction is among the earliest observed between bacteria and phytoplankton [64]. Apart from nitrogen-fixing cyanobacteria symbionts [65], described interactions between diatoms and bacteria commonly comprise members of the Rhodobacteraceae, Bacteroidota and Gammaproteobacteria [7, 66]. These interactions, however, are generally not well characterized and based on either co-culturing experiments or correlation analysis of amplicon or genomics & transcriptomics sequencing data [7, 13, 20]. This model-system centred view has shaped our general concept of the phycosphere and its best-characterized members consisting of Alteromonadaceae (Gammaproteobacteria), Flavobacteriaceae and Rhodobacteraceae, with no members of the Verrucomicrobiota or Planctomycetes being discussed as central components [67]. Herein, we didn't detect any Alteromonadaceae ipaMAGs in any phytoplankton sort, and only recovered two Rhodobacteraceae ipaMAGs from the genus *Amylibacter*, which was originally isolated from the surface of the green algae *Ulva fenestrata* [68], suggesting an association to the phycosphere in the phytoplankton cells sorted herein. Representatives of the genus have also been isolates from surface seawater [69], also indicating that they might be associated with phytoplankton, but due to the low quality of the ipaMAGs recovered herein they were not considered in downstream functional analysis. The absence of any members of Alteromonadaceae doesn't rule them out as

components of the phycosphere but suggests that they might not have been physically attached to the sorted eukaryotic cell and were therefore separated from the cell during the sorting process. Planctomycetes are often observed in high abundance within phytoplankton blooms [70, 71] and have been isolated repeatedly from the surface of macroalgae and live within their biofilm [72, 73]. They are also hypothesized to attach to particles and interact with eukaryotic phytoplankton [74] and have been seen to attach to their surface [71, 75], which is in-line with our findings, but their actual lifestyle and the nature of their interactions with phytoplankton remain elusive. Proposed characteristics involve antimicrobial functions to make them competitive despite their slow growth-rate compared to other heterotrophic bacteria, but also scavenging and algicidal activity, which has been proposed because of their correlation with phytoplankton blooming events, although genomic pathways and the actual mechanisms facilitating these functions are unknown [76, 77]. Verrucomicrobiota, like the Planctomycetota, are not part of the classical members of the phycosphere [67], although they are known to possess a diverse catalogue of polysaccharide-degrading enzymes to utilize carbohydrates produced by phytoplankton [78]. Although this suggests an association with the phycosphere, Verrucomicrobiota are not generally found in association with phytoplankton in molecular surveys of phycosphere-composition [79, 80].



**Figure 6: Functional comparison of ipaMAG lineages with reference genomes representing diverse lifestyles** (A) genome size of symbionts and selected marine reference genomes with free-living and host-dependent lifestyles. (B) Functional comparison of lineages recovered here and reference genomes, based on completeness of core metabolic pathways and accessory functions. The heatmap reflects the number of genes associated with a function or pathway. Gene counts were z-score normalized and genomes clustered hierarchically based on Manhattan distance and complete linkage clustering. Taxa names coloured in purple underlined represent ipaMAGs recovered herein.

To this end, functional profiling was performed on the ten representative high quality ipaMAGs (Table 1, Figure 5). Then, comparative genomic analyses were performed against select reference genomes from well understood cultivated bacteria with known lifestyles. These revealed a clear separation in terms of encoded functionality and metabolic pathway reduction between members of different lineages. ipaMAGs from Bacteroidota, Verrucomicrobiota and Planctomycetota clustered with reference genomes with free-living (e.g. *Prochlorococcus* [22]) or particle-attached lifestyles (e.g. *Rhodopirellula* [81]) (Figure 6), but also including lineages of pathogens with extensive free-living capacities, like *Vibrio* and *Legionella* [82, 83]. This clustering was driven by similarity in the completeness of

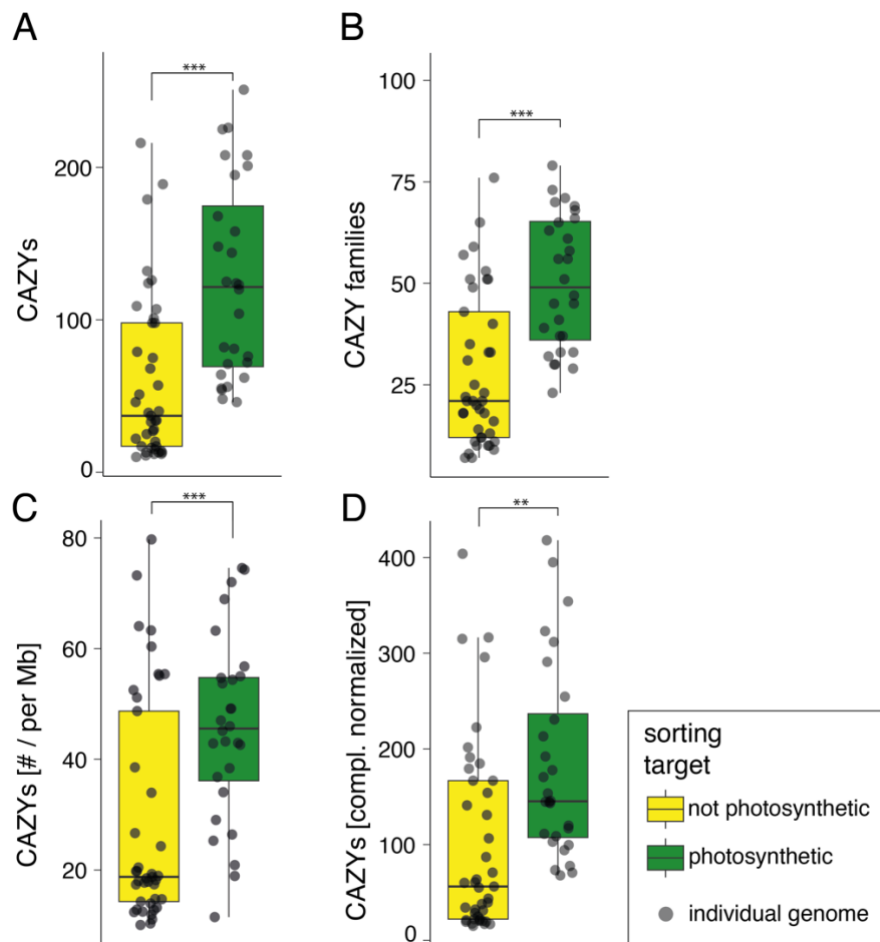
amino acid, lipid and other core metabolisms. ipaMAGs classified as Gamma- or Alphaproteobacteria, on the other hand, clustered with symbiotic reference genomes demonstrating various degrees of genome reduction and host dependency. This included organisms with a pathogenic lifestyle (*Ca. Chromulinavorax destructans* [84]) as well as representatives with a more mutualistic symbiont-host relationship (e.g. *Buchnera* [85]). This clustering according to their lifestyle was preserved even for genomes from the same lineage, with the Bacteroidota obligate intracellular symbiont *Ca. Amoebophilus asiaticus* [86] clustering with other symbionts and not with marine Bacteroidota often found in connection to phytoplankton blooms [87]. ipaMAGs from sorts representing eukaryotic phytoplankton were larger on average than those from their heterotrophic counterparts ( $4.08 \pm 1.92$  Mb vs  $2.85 \pm 2.76$  Mb estimated genome size for medium or high-quality genomes).

#### Carbohydrate-Active EnZYme diversity within phycosphere-bacteria.

A category that exhibited major Carbohydrate-Active EnZymes (CAZys), diverse families of enzymes that are generally related to the modification, synthesis, and degradation of carbohydrates or carbohydrate-modified molecules [38]. Carbohydrates are one of the major components of marine DOM, often exudated by phytoplankton during photosynthesis or during cell death. Additionally, glycosyl-groups are often added to lipids or proteins as posttranslational modifications, e.g., to protect those against degradation by other entities, impacting pathogen survival and virulence resistance [88, 89]. Carbohydrates are utilized by heterotrophic bacteria as a major carbon and energy source [90]. ipaMAGs from sorts representing eukaryotic phytoplankton encoded statistically significantly more total CAZys (Figure 7A;  $p < 0.0001$ ), more distinct families of CAZys (Figure 7B;  $p < 0.0001$ ) and more CAZys when normalized via Mb (Figure 7C) or via genome completeness (Figure 7C;  $p < 0.001$ ) than their counterparts found in association with putative heterotrophic protists.

A detailed breakdown into the CAZys encoded by the Bacteroidota, Planctomycetota and Verrucomicrobiota, the lineages we found associated with eukaryotic phytoplankton, gave insights into aspects of their potential interactions with the protists. Related ipaMAGs encoded similar CAZy families, and although there was a large overlap in the most abundant CAZys, we found distinct families missing or enriched differentially (Figure 8A). In line with previous observations, all ipaMAGs carried multiple, sometimes up to 44copies (ipaMAG-0027-1, CAJQQY01 (Bacteroidota)) of the glycosyltransferase (GT) families GT2 (especially subfamilies GT2\_2 and GT2\_3) and GT4 (Figure 8B). GTs catalyse the transfer of sugar

moiety from a donor sugar to a saccharide or non-saccharide acceptor [91]. GT2 and GT4 are large enzyme families, including various characterized enzymes, including cellulose & chitin synthases [38], but their members are polyspecific, with large variability in their substrates and functions, making homology-based assumptions into their exact role within the ipaMAGs difficult [91]. Both families are known to be abundant within the Bacteroidota lineage especially, and more generally, GT2 and GT4-family GTs comprise more than 50% of the known marine GTs [92]. In line with the unspecific GTs, all lineages encode various copies of the carbohydrate esterase (CE) families CE1 and CE4. CE1 is among the biggest and most diverse CE families, with a large variety of substrates, and CE4, similarly, catalyses de-acylation of various polysaccharides, [38].



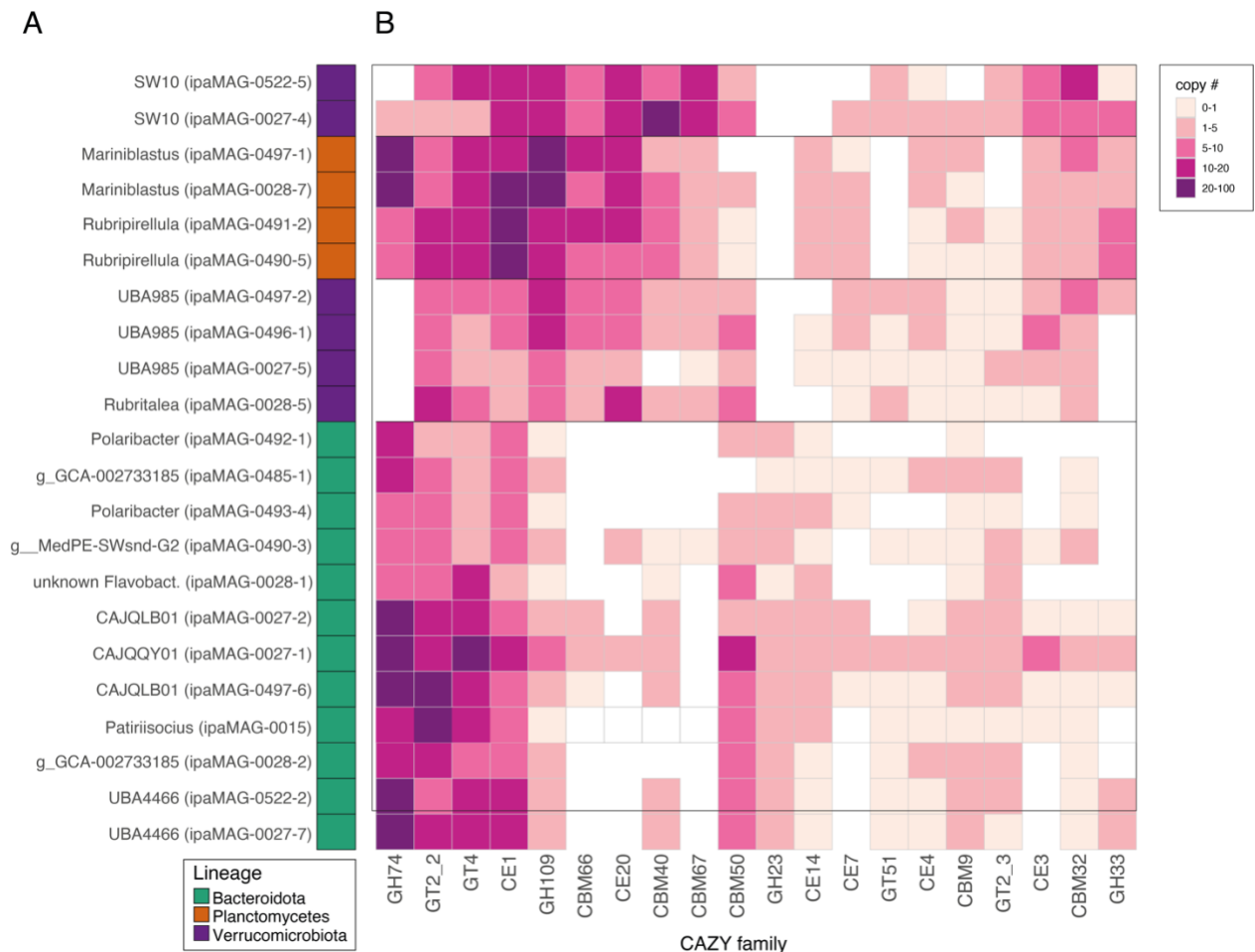
**Figure 7: CAZys in bacteria based on the associated eukaryote.** Comparisons of CAZys within bacterial SAGs recovered from assembled population and single-cell sorts, annotated based on through dbCAN3 (<https://bcb.unl.edu/dbCAN/>; CAZY db v.11). Bacterial assemblies were compared based on (A) the total number of CAZys within a SAG, (B) the number of different CAZY families, (C) the number of CAZys per Mb of assembly, and (D) normalized by the completeness of the bacterial assembly. Significance (tested through pairwise t-tests) is indicated by asterisks (\*  $p < 0.01$ ; \*\*  $p < 0.001$ ; \*\*\*  $p < 0.0001$ ).

We also identified various families of carbohydrate-binding modules (CBMs) (Figure 8B), which are generally noncatalytic and work in synteny with other CAZys. CBM66, for, example, found here primarily in Verrucomicrobiota and Planctomycetota, is linked to a range of CAZys, most commonly glycoside hydrolases (GHs) and lyases that are associated with plant cell-wall degradation, but it's specific binding properties remain unknown [93]. Throughout all lineages, we also observed the common CBMs CBM50, which binds to enzymes that cleave either chitin or peptidoglycan (incl. GH23, seen here in all Bacteroidota but missing in the other groups), and CBM9, which is found in xylanases, and has been demonstrated to bind cellulose in one case [94].

Apart from various families of CAZys whose specificity beyond carbohydrates in general remains elusive, we have been able to identify a series of CAZys abundant in the phytoplankton-associated genomes described herein that give insights into their potentially pathogenic lifestyle. Generally, the phycosphere surrounding marine phytoplankton, e.g., diatoms, has generally been considered a place of mutualistically beneficial symbiosis with the phytoplankton exuding excess DOM that is utilized by heterotrophic bacteria, which in turn provide essential nutrients and cofactors for the phytoplankton [7]. However, the extracellular proteome of diatoms also includes mucin-like proteins [95]. These proteins have a similar structure to human proteins forming a layer protecting epithelia, including the human gut, from harmful bacteria by preventing bacterial aggregation on its surface [96]. These mucin proteins remain understudied in phytoplankton, but if diatom mucins prevent the attachment of pathogenic bacteria, this group of proteins could present an important defence mechanism for the phytoplankton cell [97].

The Verrucomicrobiota and Planctomycetota ipaMAGs, but also some Bacteroidota (CAJQQY01 and g\_\_MedPE-SWsnd-G2), encode the CAZy family CE20 (Figure 8B), putatively functioning as a sialate O-acetyltransferase. We also observed GH33 (often linked with CBM40), encoding exosialidases in all three lineages. Sialic acids are often located in terminal positions in glycoproteins, glycolipids and glycans, including mucins in the human gut, and this outermost carbohydrate protects the integrity of the glycoproteins [98]. Sialic acid residues in mucin glycans for example protect the mucin glycan structure from degradation [99]. Bacteroidota present in the human gut also produce sialidases, which can promote the expansion of potential pathogens like *Salmonella* and *Escherichia coli*, who do not encode sialidases themselves [100]. Terminal sialic acid, to protect it from exosialidase (GH33) activity, is often additionally modified with O-acetyl esters [100]. However, in turn,

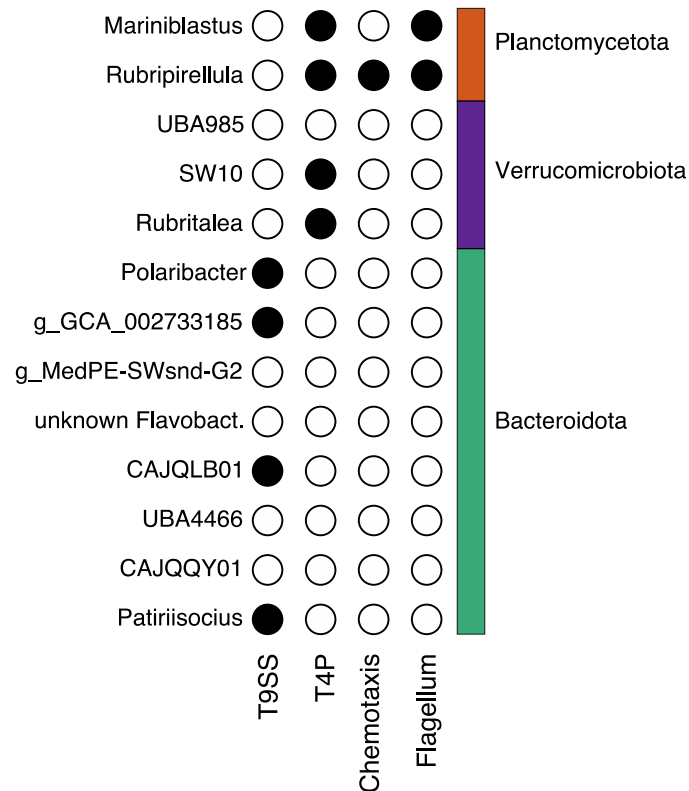
sialate O-acetyl esterases (CE20) can remove these O-acetyl esters from sialic acid, making it susceptible to cleavage through sialidases (GH33). Through the action of these two CAZy families, glycans protected with terminal sialic acid residues become exposed to further degradation through various GHs [38]. We observed both C20 and G33 within all three lineages, but especially frequently in Planctomycetota and Verrucomicrobiota.



**Figure 8: ipaMAGs recovered from phytoplankton encode more CAZys.** (A) Taxonomy and (B) top 20 CAZy families present (based on absolute copy numbers) of ipaMAGs found in association with phytoplankton, clustered hierarchically on both axis using Euclidean distance based on a distance matrix of CAZy copy numbers per family.

CE20 has also been implicated in the breakdown of cell wall xyloglucan by pathogens of vascular plants [101]. This is complemented by our detection of the family GH74 herein, which is a highly specific family of xyloglucanases [102], setting it apart from various poly-specific GH families [38]. Both CE20 and GH74 have been implicated in a pathway connected to the breakdown of xyloglucan in plants, where it acts as a barrier against

pathogens, again, through high amounts of substitutions to complicate their breakdown [101]. Xyloglucans are not encoded in all marine phytoplankton but are present in green algae [103]. For those ipaMAGs investigated more closely, we observed copies of both enzymes co-located next to each other in ipaMAG-0028-7 and observed a proximity (spanning 12 proteins) between CE20, GH74 and GH33 in ipaMAG-0027-4. The parallels in key enzymatic repertoire observed here compared to that of bacteria implicated in the breakdown of the extracellular glycoprotein layer in the human gut or vascular plant cell walls gives rise to hypothesise a potential pathogenic lifestyle for some of the ipaMAGs observed in association with phytoplankton cells herein. Given the strategy of complex modifications added to terminal glycoproteins as protective mechanisms, divergence, and high copy numbers within enzyme families to break these complex proteins could be an evolutionary mechanism shaping these CAZy families. Homologs of the same family could be specific to one or a subset of modification on, e.g., a terminal sialic acid molecule, with the repertoire of sometimes dozens of copies and a variety of enzyme families enabling the breakdown of the full variance of sialic acid modifications on a glycoprotein, making it accessible for follow-up utilization. Given that we always observed multiple bacteria associated with a single phytoplankton cell, the cumulative functional repertoire in their carbohydrate-utilizing enzymes could expand across species, with multiple lineages collaborating in this breakdown of complex cell-wall molecules.



**Figure 9: presence (black circle) and absence (white circle) of genomic systems associated with motility and cell-attachment.** All lineages with genomes of medium or high-quality were analysed. Type III or IV secretion systems were not detected in any of the queried lineages.

### Motility systems and cell surface attachment mechanisms

While harbouring larger metabolic potential across all functional categories (Figure 6), in line with their, on average, larger genomes and putatively copiotrophic lifestyle, no Planctomycetota, Bacteroidota or Verrucomicrobiota encoded a type IVA or type IVB secretion system (Figure 4), characteristic to bacteria with symbiotic or pathogenic associations with specific protists [104, 105]. We did not detect any signatures of type III secretion systems, which have previously been implicated in injecting effector proteins into the eukaryotic host cell to module late host response to bacterial attachment [106].

Multiple ipaMAG-lineages from the Bacteroidota carried a type IX secretion system (Figure 9) (T9SS), which is exclusive to the phylum [107, 108]. It is utilized by various pathogens to secrete large amounts of virulence factors which are generally attached to the outer membrane afterwards, corrupting host defence mechanisms [109], leading to its characterization as a molecular weapon against host cells. In environmental bacteria, mainly characterized isolates from soil ecosystems, T9SS genes are essential for gliding motility and

secretion of chitinase and cellulase [110]. This suggests a similar function for these marine Bacteroidota lineages with a similar lifestyle, with the T9SS likely being associated to gliding motility and export of CAZys, with a potential role in more antagonistic functions against possible host defence mechanisms in the phycosphere.

Both Planctomycetota lineages encoded more classical motility systems, with a complete flagellum in both, and the potential for chemotaxis in the *Rubripirellula* (Figure 9). Both lineages, and two out of three Verrucomicrobiota lineages also included a type 4 pilus, dynamic fibers with extensive functional diversity spanning virulence and protein secretion as well as biofilm formation, aggregation, surface sensing and motility [111]. ipaMAG-0491-2 (*Rubripirellula*, Planctomycetota) encoded a Tad pilus system, a subtype of the type IV pilus (T4P) nanomachine [111, 112]. Its origin lays in an archaeal T4P system, explaining its distinction from other T4P subtypes, but it horizontally moved into bacteria in an interkingdom transfer, and has since extensively spread within them [113]. The functionality of Tad pili has not been characterized in most bacteria that possess them, but initial studies in cultured model systems indicate a significant functional diversity [112]. Tad systems are prevalent in pathogens and commensal symbionts of mammals and plants, and Tad has been implicated in host-pathogen interactions and pathogenesis [114, 115]. Characterized Tad systems, among others, play an essential role in surface attachment and aggregation in *Aggregatibacter actinomycetemcomitans* [116], guide surface sensing and attachment in *Caulobacter crescentus* [117], allows twitching motility in *Liberibacter crescens* [118], enable *Bifidobacterium breve* adherence to epithelia [119], and trigger contact-dependent killing of prey cells in *Myxococcus xanthus* [120]. Although the wide array of functionality of the Tad pilus does not allow a definite prediction of its exact function in ipaMAG-0491-2 (*Rubripirellula*), its predominant role in attachment to various surfaces suggests that it enables *Rubripirellula* to attach to the phytoplankton cell to stay within the phycosphere. This is further supported by the observation that the Tad pilus is the defining feature of putative Rhodobacterales phycosphere colonizers of diatom hosts [13].

Interestingly, we do not see various lineages of bacteria regularly associated with the phycosphere in the literature, for example Gammaproteobacteria groups Oceanospirillales or Alteromonadaceae [67, 121]. Both lineages are usually omnipresent in the surface ocean but it's theoretically possible that phycosphere-associated representatives were not present and associated with the phytoplankton cells at the time of sorting. Alternatively, and arguably more likely, these lineages represent phycosphere-associated bacteria that do not physically attach to the eukaryotic phytoplankton and are therefore not co-sorted with the protists during

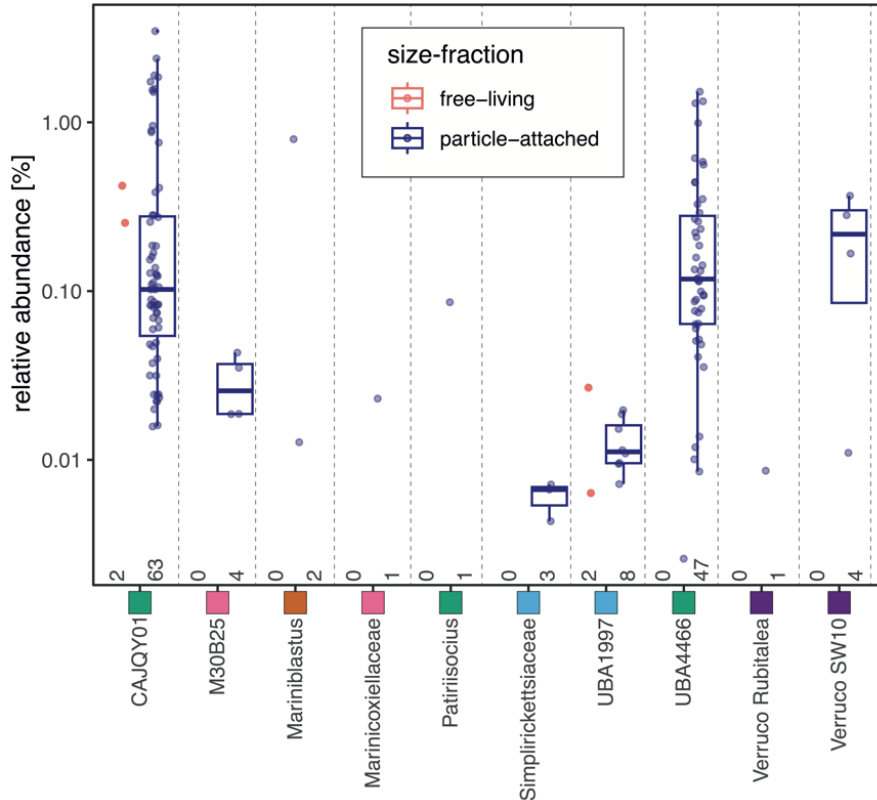
sampling. They could employ a different strategy, not attaching to the cell but being more motile and able to conduct chemotaxis to find and stay with the phycosphere instead, a common approach previously observed in various marine bacteria [67].

Cumulatively, we observed systems associated with motility, surface attachment and sometimes virulence in eight out of 13 lineages analysed, but due to some lineages not having complete ipaMAGs, systems could have been missed in those lineages. Generally, mobility provides an advantage to cells that desire to find the phycosphere of a phytoplankton cell quicker, because it greatly increases the encounter rate with such a microenvironment [67]. This encounter rate can be boosted further with the ability for chemotaxis (here observed only in the *Rubripirellula*) which provides the ability to ‘sense’ gradients in DOM, allowing the cell to use them as a ‘map’ towards the phycosphere more directly, instead of moving uncoordinated and with no direction [67, 122]. Similarly, being able to attach to the phytoplankton cell via, e.g., a pilus or secretion system can be an advantage because it conserves the energy that is required for continuous chemotaxis and motility towards the phytoplankton cell [123, 124], a mechanism observed in the *Roseobacter* group (Alteromonadaceae) associated with the diatom *Asterionellopsis glacialis* [125]. However, a flagellar machinery or chemotaxis require extensive genomic and energy resources, leading to a trade-off between the benefits of finding and getting to the phycosphere faster vs spending a lot of resources in doing so [67]. This opens the possibility that some lineages, for example the UBA4466 and CAJQQY01 described here, could opt for lower encounter rates, but do so in an energy-efficient way as an alternative evolutionary trajectory.

#### Global distribution and particle-attachment of bacteria associated with marine protists

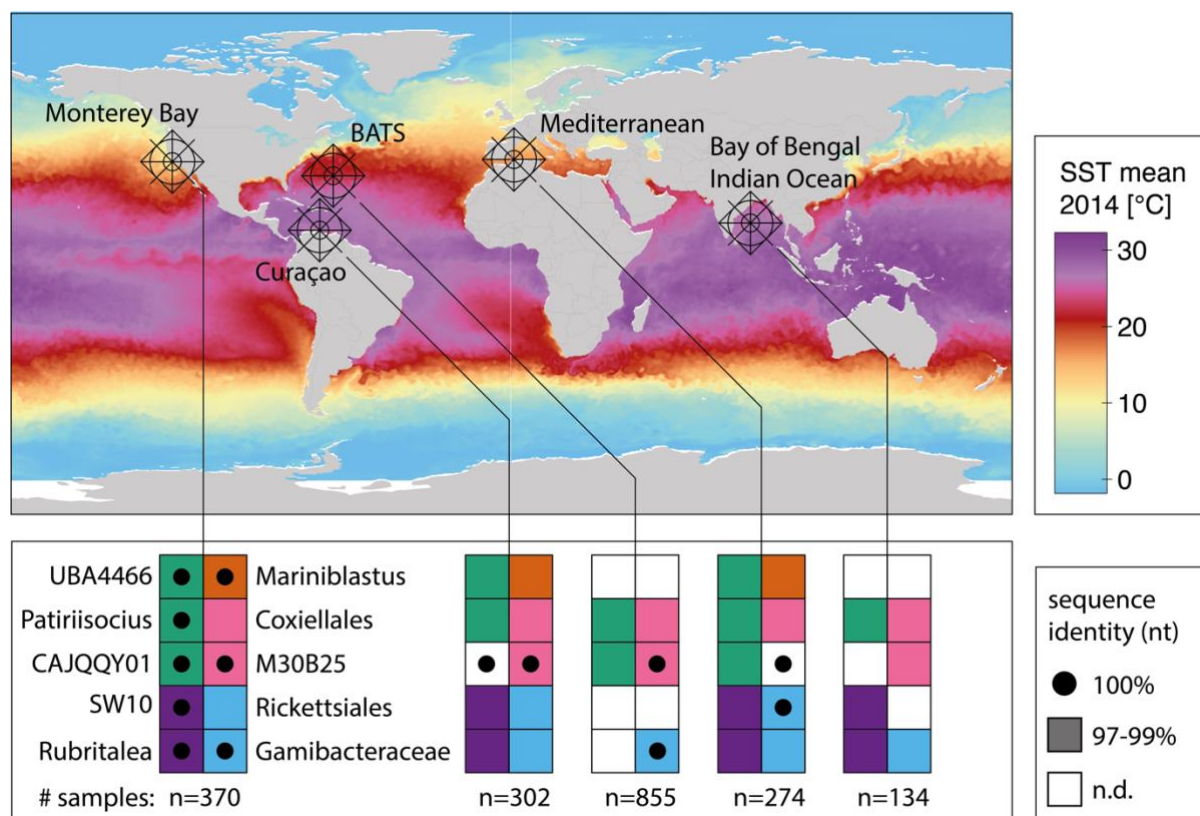
Comparisons of 16S rRNA gene sequences from the ten high-quality ipaMAG lineages with size-fractionated sequences showed that they are rarely observed in the free-living size fraction (0.2-1.2µm), but all, although some are rare, were observed in the particle-attached (>5µm) fraction (Figure 5). This supports our finding and suggests that these lineages indeed all are physically connected to their eukaryotic cells. Furthermore, it suggests that either none or only a small subset of the population spends time in the free-living fraction, i.e., not attached to another cell. When spending a vast majority of their time attached directly to the surface of a phytoplankton cell, this essentially renders a motility system useless and a waste of resources, potentially explaining why we don’t see flagella and chemotaxis systems widespread within the Planctomycetota, Bacteroidota and Verrucomicrobiota. The rarity of

the lineages found with heterotrophic protists, on the other side, compare to dominant lineages like *Prochlorococcus*, as well as their almost exclusively particle attached lifestyle also both support the hypothesis, that these lineages do not represent prey organisms.



**Figure 10: ipaMAGs represent particle-associated bacteria.** Relative abundance (95% ID cutoff to query sequence) was only visualised for values above 0 for a total of 432 samples as in Comstock et al. [126]. Boxplots for lineage and size-fractions were only shown when the lineage was observed in at least three samples.

To verify that the protist-associated lineages we recovered herein were not endemic to Monterey Bay, where most of the sorting took place, we surveyed a series of in-depth amplicon studies covering multiple ocean basins. All lineages showed global distribution patterns spanning coastal and open ocean ecosystems (Figure 11). This indicates that these lineages generally do not represent endemic lineages from within Monterey Bay, where most of them were recovered, or a specific niche of marine ecosystems, but instead highlights a larger relevance of these lineages for global marine ecosystems.



**Figure 11: Global distribution of ipaMAG lineages.** Subtropical and tropical sites searched for presence of 16S rRNA gene sequences from ten high-quality ipaMAG lineages (top panel); the background color gradient shows mean 2014 sea surface temperature (SST) based on data obtained from NOAA. Amplicon data comes from published studies with differing sample collection intensity and years ranging from 2009 to 2019. Presence of ten bacterial lineages physically co-associated with protists across sites determined by detection of amplicons within subtropical and tropical oceans. Data is shown only to taxa for which high-quality ipaMAGs were recovered. The number of samples searched per dataset is indicated below the respective panel.

## Conclusions

The recovery of ipaMAGs from heterotrophs provided an important comparator for highlighting differences in the bacteria that associate physically with phytoplankton versus heterotrophic protists. In the following chapter, data from the heterotrophic protists and their associated bacteria is used to more deeply examine ecological and evolutionary aspects of bacteria affiliated with key predatory protists. From an oceanographic perspective, these microbial interactions seem to effect protists and their abundance in major aspects. Some members of the phycosphere could be competing, or even feeding on phytoplankton, regulating population size, while others could be beneficial, providing nutrients, cofactors (such as vitamins). Similarly, intracellular symbionts of heterotrophic protists might provide cofactors benefiting the predator and providing supplementary compounds lacking from their nutrition. Energy parasitism might on the other hand harm the predator, thus also affecting

the abundance and success of heterotrophic protists, which in turn would change their impact on the food web and movement of carbon through the microbial community. A lot of these potential effects remain abstract, due to the lack of insights into the exact role of various players within the phycosphere and the almost complete lack of any microbiome of heterotrophic protists, asking for more culture-independent, high-resolution work on interactions between protists and bacteria.

Our current insights into the dynamics and interactions between the phytoplankton “host” cell and its associated bacteria within the phycosphere are derived from culture-based studies [7]. The concept of the phycosphere as a place of widespread mutualistic co-association is driven by the benefits of co-culturing of phytoplankton with bacteria frequently observed in the laboratory [127, 128]. Although certainly true to some degree, this raises the question if this understanding of phycosphere as beneficial might be affected by a co-culture-driven bias against antagonistic (algicidal) bacteria during the isolation process. While the phycosphere can be seen as a place for a beneficial symbiosis for both sides, with phytoplankton receiving cofactors and access to micronutrients in return for exuded DOM, this is contradicted by the widespread competition of marine bacteria and phytoplankton for inorganic nutrients [67]. The existence of marine bacteria with algicidal activity is well established although they have not been discussed in the context of the phycosphere historically, but more generally in their role in terminating algal blooms [129, 130], a separation that has been overcome only recently [130]. This poses the question if bacteria in the phycosphere solely depend passively on the DOM produced and exuded by phytoplankton, or if they possess and could actively use their biochemical potential to utilize the particulate living algal cell [129], potentially damaging, or even killing it in the process? We still lack extensive insights into these bacteria and how widespread this bacterial lifestyle is, but two Bacteroidota (*Kordia algicida*) [131] and *Croceibacter atlanticus* [132] have been characterized as algicidal bacteria killing diatoms, indicating that some members of the phycosphere are antagonists. The lifestyle of these bacteria can, to a certain degree, depend on the algal cell they are associated with, supported by the observation that algicidal bacteria can act beneficial members of the phycosphere of one and pathogenic when associated with their non-native host [133].

The approach employed here allowed investigation of interactions that occur in the wild and circumvented any biases that might arise from cultivation. The results showed that various bacterial members of the phycosphere likely encode an extensive biochemical machinery that can utilize components beyond exuded DOM. This includes members of the Bacteroidota, a group that comprises classical model systems of algicidal bacteria like *Kordia*

*algicida*, but also Verrucomicrobiota and Planctomycetota lineages, both of which are generally not considered classical groups of algicidal bacteria. Focusing on the CAZys encoded by the associated bacterial community, we characterized various potential pathways and enzymatic cascades that could degrade key components of the intact phytoplankton cell-wall in all lineages recovered. These lineages could also encode functions beneficial to their associated phytoplankton, and their potential algicidal activity could be restricted to specific hosts or environmental conditions. It will require additional studies, particularly those focussing on expression of pathways, to understand if these enzymes are continuously expressed, attacking the live phytoplankton cell, potentially counteracted by continuous cell-wall repair by the phytoplankton cell, or if these pathways are triggered by environmental conditions, for example competition for inorganic nutrients or the availability of more easily accessible DOM (or lack thereof) in the phycosphere.

**Author Contributions**

F.W. and A.Z.W. conceptualized the analysis. C.P. and A.Z.W. designed the sampling strategy and conducted single-cell sorting. L.S. and R.M. prepared single-cell sequencing plates and generated sequencing data. F.W. performed bioinformatics analysis and wrote the manuscript. A.Z.W. provided edits to the manuscript.

## References

1. Field CB. Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science* (1979) 1998; **281**: 237–240.
2. Thornton DCO. Dissolved organic matter (DOM) release by phytoplankton in the contemporary and future ocean. *Eur J Phycol* 2014; **49**: 20–46.
3. Lancelot C. Extracellular release of small and large molecules by phytoplankton in the Southern Bight of the North Sea. *Estuar Coast Shelf Sci* 1984; **18**: 65–77.
4. Bell W, Mitchell R. Chemotactic and growth responses of marine bacteria to algal extracellular products. *Biol Bull* 1972; **143**: 265–277.
5. Seymour JR, Ahmed T, Durham WM, Stocker R. Chemotactic response of marine bacteria to the extracellular products of *Synechococcus* and *Prochlorococcus*. *Aquatic Microbial Ecology* 2010; **59**: 161–168.
6. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B. The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res* 2009; **37**: D233–D238.
7. Helliwell KE, Shibl AA, Amin SA. The Diatom Microbiome: New Perspectives for Diatom-Bacteria Symbioses. In: Falciatore A, Mock T (eds). *The Molecular Life of Diatoms*. 2022. Springer International Publishing, Cham, pp 679–712.
8. Nagata T. Production mechanisms of dissolved organic matter. *Microbial ecology of the oceans* 2000; 121–152.
9. Pete R, Davidson K, Hart MC, Gutierrez T, Miller AEJ. Diatom derived dissolved organic matter as a driver of bacterial productivity: The role of nutrient limitation. *J Exp Mar Biol Ecol* 2010; **391**: 20–26.
10. Mykkestad SM. Release of extracellular products by phytoplankton with special emphasis on polysaccharides. *Science of The Total Environment* 1995; **165**: 155–164.
11. Behringer G, Ochsenkühn MA, Fei C, Fanning J, Koester JA, Amin SA. Bacterial communities of diatoms display strong conservation across strains and time. *Front Microbiol* 2018; **9**: 659.
12. Cooper MB, Kazamia E, Helliwell KE, Kudahl UJ, Sayer A, Wheeler GL, et al. Cross-exchange of B-vitamins underpins a mutualistic interaction between *Ostreococcus tauri* and *Dinoroseobacter shibae*. *ISME J* 2019; **13**: 334–345.
13. Isaac A, Francis B, Amann RI, Amin SA. Tight adherence (Tad) pilus genes indicate putative niche differentiation in phytoplankton bloom associated Rhodobacterales. *Front Microbiol* 2021; **12**: 718297.

14. Moran MA, Kujawinski EB, Schroer WF, Amin SA, Bates NR, Bertrand EM, et al. Microbial metabolites in the marine carbon cycle. *Nat Microbiol* 2022; **7**: 508–523.
15. Decelle J, Carradec Q, Pochon X, Henry N, Romac S, Mahé F, et al. Worldwide occurrence and activity of the reef-building coral symbiont Symbiodinium in the open ocean. *Current Biology* 2018; **28**: 3625–3633.
16. Balzano S, Corre E, Decelle J, Sierra R, Wincker P, Da Silva C, et al. Transcriptome analyses to investigate symbiotic relationships between marine protists. *Front Microbiol* 2015; **6**: 98.
17. Keeling PJ. The endosymbiotic origin, diversification and fate of plastids. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2010; **365**: 729–748.
18. Wilken S, Yung CCM, Poirier C, Massana R, Jimenez V, Worden AZ. Choanoflagellates alongside diverse uncultured predatory protists consume the abundant open-ocean cyanobacterium Prochlorococcus. *Proceedings of the National Academy of Sciences* 2023; **120**: e2302388120.
19. Zubkov M V, Tarran GA. High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature* 2008; **455**: 224–226.
20. Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, et al. Determinants of community structure in the global plankton interactome. *Science (1979)* 2015; **348**: 1262073.
21. Karlusich JJP, Ibarbalz FM, Bowler C. Exploration of marine phytoplankton: from their historical appreciation to the omics era. *J Plankton Res* 2020.
22. Chisholm SW, Olson RJ, Zettler ER, Goericke R, Waterbury JB, Welschmeyer NA. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* 1988; **334**: 340–343.
23. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**: 357–359.
24. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.
25. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; **25**: 1043–1055.
26. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 2021; **6**: 3–6.
27. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences* 2019; **116**: 20574 LP – 20583.

28. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015; **12**: 59–60.
29. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2020; **36**: 1925–1927.
30. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics* 2010; **11**: 119.
31. Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, et al. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nat Biotechnol* 2017; **35**: 725–731.
32. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. 2016. PeerJ Preprints.
33. Huerta-Cepas J, Forslund K, Coelho LP, Szklarczyk D, Jensen LJ, von Mering C, et al. Fast Genome-Wide Functional Annotation through Orthology Assignment by eggNOG-Mapper. *Mol Biol Evol* 2017; **34**: 2115–2122.
34. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019; **47**: D309–D314.
35. Mistry J, Chuguransky S, Williams L, Qureshi M, Salazar GA, Sonnhammer ELL, et al. Pfam: The protein families database in 2021. *Nucleic Acids Res* 2021; **49**: D412–D419.
36. Finn RD, Clements J, Eddy SR. HMMER web server: interactive sequence similarity searching. *Nucleic Acids Res* 2011; **39**: W29–W37.
37. Abby SS, Néron B, Ménager H, Touchon M, Rocha EPC. MacSyFinder: A Program to Mine Genomes for Molecular Systems with an Application to CRISPR-Cas Systems. *PLoS One* 2014; **9**: e110726-.
38. Drula E, Garron M-L, Dogan S, Lombard V, Henrissat B, Terrapon N. The carbohydrate-active enzyme database: functions and literature. *Nucleic Acids Res* 2022; **50**: D571–D577.
39. Kanehisa M, Sato Y, Morishima K. BlastKOALA and GhostKOALA: KEGG Tools for Functional Characterization of Genome and Metagenome Sequences. *J Mol Biol* 2016; **428**: 726–731.
40. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; **25**: 1972–1973.
41. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 2010; **5**: e9490-.

42. Gruber-Vodicka HR, B SBK, Elmar P. phyloFlash: Rapid Small-Subunit rRNA Profiling and Targeted Assembly from Metagenomes. *mSystems* 2020; **5**: e00920-20.
43. Pruesse E, Peplies J, Glöckner FO. SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics* 2012; **28**: 1823–1829.
44. Barbera P, Kozlov AM, Czech L, Morel B, Darriba D, Flouri T, et al. EPA-ng: Massively Parallel Evolutionary Placement of Genetic Sequences. *Syst Biol* 2019; **68**: 365–369.
45. Czech L, Barbera P, Stamatakis A. Genesis and Gappa: Processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* 2020; **36**: 3263–3265.
46. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017; **8**: 28–36.
47. Eckmann CA, Bachy C, Wittmers F, Strauss J, Blanco-Bercial L, Vergin KL, et al. Recurring seasonality exposes dominant species and niche partitioning strategies of open ocean picoeukaryotic algae. *Commun Earth Environ* 2024; **5**: 266.
48. Eckmann CA, Eberle JS, Wittmers F, Wilken S, Bergauer K, Poirier C, et al. Eukaryotic algal community composition in tropical environments from solar salterns to the open sea. *Front Mar Sci* 2023; **10**.
49. Freilich MA, Poirier C, Dever M, Alou-Font E, Allen J, Cabornero A, et al. 3D intrusions transport active surface microbial assemblages to the dark ocean. *Proceedings of the National Academy of Sciences* 2024; **121**: e2319937121.
50. Strauss J, Choi CJ, Grone J, Wittmers F, Jimenez V, Makareviciute-Fichtner K, et al. The Bay of Bengal exposes abundant photosynthetic picoplankton and newfound diversity along salinity-driven gradients. *Environ Microbiol* 2023; **n/a**.
51. Grone J, Poirier C, Abbott K, Wittmers F, Jaeger GS, Mahadevan A, et al. A single Prochlorococcus ecotype dominates the tropical Bay of Bengal with ultradian growth. *Environ Microbiol* 2024; **26**: e16605.
52. Needham DM, Poirier C, Bachy C, George EE, Wilken S, Yung CCM, et al. The microbiome of a bacterivorous marine choanoflagellate contains a resource-demanding obligate bacterial associate. *Nat Microbiol* 2022; **7**: 1466–1479.
53. Harbeitner RC, Wittmers F, Yung CCM, Eckmann CA, Hehenberger E, Blum M, et al. Gradients of bacteria in the oceanic water column reveal finely-resolved vertical distributions. *PLoS One* 2024; **19**: e0298139-.
54. Laslett D, Canback B. ARAGORN, a program to detect tRNA genes and tmRNA genes in nucleotide sequences. *Nucleic Acids Res* 2004; **32**: 11–16.

55. Keeling PJ, Burki F. Progress towards the Tree of Eukaryotes. *Current Biology* 2019; **29**: R808–R817.
56. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, et al. Charting the complexity of the marine microbiome through single-cell genomics. *Cell* 2019; **179**: 1623–1635.e11.
57. Giovannoni SJ, Cameron Thrash J, Temperton B. Implications of streamlining theory for microbial ecology. *ISME J* 2014; **8**: 1553–1565.
58. Grote J, Thrash JC, Huggett MJ, Landry ZC, Carini P, Giovannoni SJ, et al. Streamlining and Core Genome Conservation among Highly Divergent Members of the SAR11 Clade. *mBio* 2012; **3**: e00252-12.
59. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 2012; **10**: 13–26.
60. Bentley SD, Parkhill J. Comparative Genomic Structure of Prokaryotes. *Annu Rev Genet* 2004; **38**: 771–791.
61. Moran NA. Microbial minimalism: genome reduction in bacterial pathogens. *Cell* 2002; **108**: 583–586.
62. Nayfach S, Roux S, Seshadri R, Udwy D, Varghese N, Schulz F, et al. A genomic catalog of Earth’s microbiomes. *Nat Biotechnol* 2021; **39**: 499–509.
63. Tully BJ, Graham ED, Heidelberg JF. The reconstruction of 2,631 draft metagenome-assembled genomes from the global oceans. *Sci Data* 2018; **5**: 170203.
64. Jolley ET, Jones AK. The interaction between *Navicula muralis* Grunow and an associated species of *Flavobacterium*. *British phycollogical journal* 1977; **12**: 315–328.
65. Foster RA, Kuypers MMM, Vagner T, Paerl RW, Musat N, Zehr JP. Nitrogen fixation and transfer in open ocean diatom–cyanobacterial symbioses. *ISME J* 2011; **5**: 1484–1493.
66. Baker LJ, Kemp PF. Exploring bacteria–diatom associations using single-cell whole genome amplification. *Aquatic Microbial Ecology* 2014; **72**: 73–88.
67. Seymour JR, Amin SA, Raina J-B, Stocker R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nat Microbiol* 2017; **2**: 17065.
68. Nedashkovskaya OI, Kukhlevskiy AD, Zhukova N V, Kim SB. *Amylibacter ulvae* sp. nov., a new alphaproteobacterium isolated from the Pacific green alga *Ulva fenestrata*. *Arch Microbiol* 2016; **198**: 251–256.
69. Teramoto M, Nishijima M. *Amylibacter marinus* gen. nov., sp. nov., isolated from surface seawater. *Int J Syst Evol Microbiol* 2014; **64**: 4016–4020.

70. Morris RM, Longnecker K, Giovannoni SJ. *Pirellula* and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. *Environ Microbiol* 2006; **8**: 1361–1370.
71. Cai H-Y, Yan Z, Wang A-J, Krumholz LR, Jiang H-L. Analysis of the Attached Microbial Community on Mucilaginous Cyanobacterial Aggregates in the Eutrophic Lake Taihu Reveals the Importance of Planctomycetes. *Microb Ecol* 2013; **66**: 73–83.
72. Faria M, Bordin N, Kizina J, Harder J, Devos D, Lage OM. Planctomycetes attached to algal surfaces: Insight into their genomes. *Genomics* 2018; **110**: 231–238.
73. Ivanova AA, Philippov DA, Kulichevskaya IS, Dedysn SN. Distinct diversity patterns of Planctomycetes associated with the freshwater macrophyte *Nuphar lutea* (L.) Smith. *Antonie Van Leeuwenhoek* 2018; **111**: 811–823.
74. Wiegand S, Jogler M, Jogler C. On the maverick Planctomycetes. *FEMS Microbiol Rev* 2018; **42**: 739–760.
75. Kimbrel JA, Samo TJ, Ward C, Nilson D, Thelen MP, Siccardi A, et al. Host selection and stochastic effects influence bacterial community assembly on the microalgal phycosphere. *Algal Res* 2019; **40**: 101489.
76. Jeske O, Surup F, Ketteniß M, Rast P, Förster B, Jogler M, et al. Developing techniques for the utilization of Planctomycetes as producers of bioactive molecules. *Front Microbiol* 2016; **7**: 205668.
77. Pizzetti I, Gobet A, Fuchs BM, Amann R, Fazi S. Abundance and diversity of Planctomycetes in a Tyrrhenian coastal system of central Italy. *Aquatic Microbial Ecology* 2011; **65**: 129–141.
78. Sichert A, Corzett CH, Schechter MS, Unfried F, Markert S, Becher D, et al. Verrucomicrobia use hundreds of enzymes to digest the algal polysaccharide fucoidan. *Nat Microbiol* 2020; **5**: 1026–1039.
79. Ramanan R, Kang Z, Kim B-H, Cho D-H, Jin L, Oh H-M, et al. Phycosphere bacterial diversity in green algae reveals an apparent similarity across habitats. *Algal Res* 2015; **8**: 140–144.
80. Câmara dos Reis M, Romac S, Le Gall F, Marie D, Frada MJ, Koplovitz G, et al. Exploring the phycosphere of *Emiliana huxleyi*: From bloom dynamics to microbiome assembly experiments. *Mol Ecol* 2023; **32**: 6507–6522.
81. Kallscheuer N, Jogler M, Wiegand S, Peeters SH, Heuer A, Boedeker C, et al. Three novel *Rubripirellula* species isolated from plastic particles submerged in the Baltic Sea and the estuary of the river Warnow in northern Germany. *Antonie Van Leeuwenhoek* 2020; **113**: 1767–1778.

82. Steinert M, Hentschel U, Hacker J. *Legionella pneumophila*: an aquatic microbe goes astray. *FEMS Microbiol Rev* 2002; **26**: 149–162.
83. Ramamurthy T, Chowdhury G, Pazhani GP, Shinoda S. *Vibrio fluvialis*: an emerging human pathogen. *Front Microbiol* 2014; **5**.
84. Deeg CM, Zimmer MM, George EE, Husnik F, Keeling PJ, Suttle CA. Chromulinavorax destructans, a pathogen of microzooplankton that provides a window into the enigmatic candidate phylum Dependotia. *PLoS Pathog* 2019; **15**: e1007801-
85. Douglas AE. Nutritional Interactions in Insect-Microbial Symbioses: Aphids and Their Symbiotic Bacteria Buchnera. *Annu Rev Entomol* 1998; **43**: 17–37.
86. Schmitz-Esser S, Patrick T, Roland A, Jacqueline M, Michael W, Thomas R, et al. The Genome of the Amoeba Symbiont “Candidatus Amoebophilus asiaticus” Reveals Common Mechanisms for Host Cell Interaction among Amoeba-Associated Bacteria. *J Bacteriol* 2010; **192**: 1045–1057.
87. Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM, et al. Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science (1979)* 2012; **336**: 608–611.
88. Limoli DH, Jones CJ, Wozniak DJ. Bacterial extracellular polysaccharides in biofilm formation and function. *Microbial Biofilms* 2015; 223–247.
89. Cress BF, Englaender JA, He W, Kasper D, Linhardt RJ, Koffas MAG. Masquerading microbial pathogens: capsular polysaccharides mimic host-tissue molecules. *FEMS Microbiol Rev* 2014; **38**: 660–697.
90. Mühlenbruch M, Grossart H-P, Eigemann F, Voss M. Mini-review: Phytoplankton-derived polysaccharides in the marine environment and their interactions with heterotrophic bacteria. *Environ Microbiol* 2018; **20**: 2671–2685.
91. Breton C, Šnajdrová L, Jeanneau C, Koča J, Imberty A. Structures and mechanisms of glycosyltransferases. *Glycobiology* 2006; **16**: 29R-37R.
92. Chen B, Liu G, Chen Q, Wang H, Liu L, Tang K. Discovery of a novel marine Bacteroidetes with a rich repertoire of carbohydrate-active enzymes. *Comput Struct Biotechnol J* 2024; **23**: 406–416.
93. Cuskin F, Flint JE, Gloster TM, Morland C, Baslé A, Henrissat B, et al. How nature can exploit nonspecific catalytic and carbohydrate binding modules to create enzymatic specificity. *Proceedings of the National Academy of Sciences* 2012; **109**: 20889–20894.

94. Boraston AB, Creagh AL, Alam MM, Kormos JM, Tomme P, Haynes CA, et al. Binding specificity and thermodynamics of a family 9 carbohydrate-binding module from *Thermotoga maritima* xylanase 10A. *Biochemistry* 2001; **40**: 6240–6247.
95. Buhmann MT, Schulze B, Förderer A, Schleheck D, Kroth PG. Bacteria may induce the secretion of mucin-like proteins by the diatom *Phaeodactylum tricornutum*. *J Phycol* 2016; **52**: 463–474.
96. Caldara M, Friedlander RS, Kavanaugh NL, Aizenberg J, Foster KR, Ribbeck K. Mucin biopolymers prevent bacterial aggregation by retaining cells in the free-swimming state. *Current Biology* 2012; **22**: 2325–2330.
97. Poulson-Ellestad KL, Jones CM, Roy J, Viant MR, Fernández FM, Kubanek J, et al. Metabolomics and proteomics reveal impacts of chemically mediated competition on marine plankton. *Proceedings of the National Academy of Sciences* 2014; **111**: 9009–9014.
98. Tailford LE, Crost EH, Kavanaugh D, Juge N. Mucin glycan foraging in the human gut microbiome. *Front Genet* 2015; **6**: 122393.
99. Glover JS, Ticer TD, Engevik MA. Characterizing the mucin-degrading capacity of the human gut microbiota. *Sci Rep* 2022; **12**: 8456.
100. Robinson LS, Lewis WG, Lewis AL. The sialate O-acetyltransferase EstA from gut Bacteroidetes species enables sialidase-mediated cross-species foraging of 9-O-acetylated sialoglycans. *Journal of Biological Chemistry* 2017; **292**: 11861–11872.
101. Vieira PS, Bonfim IM, Araujo EA, Melo RR, Lima AR, Fessel MR, et al. Xyloglucan processing machinery in *Xanthomonas* pathogens and its role in the transcriptional activation of virulence factors. *Nat Commun* 2021; **12**: 4049.
102. Arnal G, Stogios PJ, Asohan J, Attia MA, Skarina T, Viborg AH, et al. Substrate specificity, regiospecificity, and processivity in glycoside hydrolase family 74. *Journal of Biological Chemistry* 2019; **294**: 13233–13247.
103. Ikegaya H, Hayashi T, Kaku T, Iwata K, Sonobe S, Shimmen T. Presence of xyloglucan-like polysaccharide in *Spirogyra* and possible involvement in cell–cell attachment. *Phycological Res* 2008; **56**: 216–222.
104. Low HH, Gubellini F, Rivera-Calzada A, Braun N, Connery S, Dujeancourt A, et al. Structure of a type IV secretion system. *Nature* 2014; **508**: 550–553.
105. Souza DP, Oka GU, Alvarez-Martinez CE, Bisson-Filho AW, Dunger G, Hobeika L, et al. Bacterial killing via a type IV secretion system. *Nat Commun* 2015; **6**: 6453.
106. Kay S, Bonas U. How *Xanthomonas* type III effectors manipulate the host plant. *Curr Opin Microbiol* 2009; **12**: 37–43.

107. Gorasia DG, Veith PD, Reynolds EC. The type IX secretion system: advances in structure, function and organisation. *Microorganisms* 2020; **8**: 1173.
108. Sato K, Sakai E, Veith PD, Shoji M, Kikuchi Y, Yukitake H, et al. Identification of a new membrane-associated protein that influences transport/maturation of gingipains and adhesins of *Porphyromonas gingivalis*. *Journal of Biological Chemistry* 2005; **280**: 8668–8677.
109. Lasica AM, Ksiazek M, Madej M, Potempa J. The type IX secretion system (T9SS): highlights and recent insights into its structure and function. *Front Cell Infect Microbiol* 2017; **7**: 215.
110. S KS, J MM. *Flavobacterium johnsoniae* Chitinase ChiA Is Required for Chitin Utilization and Is Secreted by the Type IX Secretion System. *J Bacteriol* 2014; **196**: 961–970.
111. Ellison CK, Whitfield GB, Brun Y V. Type IV Pili: dynamic bacterial nanomachines. *FEMS Microbiol Rev* 2022; **46**: fuab053.
112. Whitfield GB, Brun Y V. The type IVc pilus: just a Tad different. *Curr Opin Microbiol* 2024; **79**: 102468.
113. Denise R, Abby SS, Rocha EPC. Diversification of the type IV filament superfamily into machines for adhesion, protein secretion, DNA uptake, and motility. *PLoS Biol* 2019; **17**: e3000390-.
114. Pu M, Duriez P, Arazi M, Rowe-Magnus DA. A conserved tad pilus promotes *Vibrio vulnificus* oyster colonization. *Environ Microbiol* 2018; **20**: 828–841.
115. Clock SA, Planet PJ, Perez BA, Figurski DH. Outer membrane components of the Tad (tight adherence) secretion of *Aggregatibacter actinomycetemcomitans*. *J Bacteriol* 2008; **190**: 980–990.
116. Tomich M, Planet PJ, Figurski DH. The tad locus: postcards from the widespread colonization island. *Nat Rev Microbiol* 2007; **5**: 363–375.
117. Ellison CK, Kan J, Dillard RS, Kysela DT, Ducret A, Berne C, et al. Obstruction of pilus retraction stimulates bacterial surface sensing. *Science (1979)* 2017; **358**: 535–538.
118. Cai L, Jain M, Sena-Vélez M, Jones KM, Fleites LA, Heck M, et al. Tad pilus-mediated twitching motility is essential for DNA uptake and survival of *Liberibacter*s. *PLoS One* 2021; **16**: e0258583-.
119. O'Connell Motherway M, Houston A, O'Callaghan G, Reunanen J, O'Brien F, O'Driscoll T, et al. A *Bifidobacterium* pilus-associated protein promotes colonic epithelial proliferation. *Mol Microbiol* 2019; **111**: 287–301.

120. Seef S, Herrou J, de Boissier P, My L, Brasseur G, Robert D, et al. A Tad-like apparatus is required for contact-dependent prey killing in predatory social bacteria. *Elife* 2021; **10**: e72409.
121. Crenn K, Duffieux D, Jeanthon C. Bacterial epibiotic communities of ubiquitous and abundant marine diatoms are distinct in short-and long-term associations. *Front Microbiol* 2018; **9**: 368508.
122. Stocker R. Marine Microbes See a Sea of Gradients. *Science (1979)* 2012; **338**: 628 LP – 633.
123. Smriga S, Fernandez VI, Mitchell JG, Stocker R. Chemotaxis toward phytoplankton drives organic matter partitioning among marine bacteria. *Proceedings of the National Academy of Sciences* 2016; **113**: 1576–1581.
124. Taylor JR, Stocker R. Trade-Offs of Chemotactic Foraging in Turbulent Water. *Science (1979)* 2012; **338**: 675–679.
125. Fei C, Ochsenkühn MA, Shibl AA, Isaac A, Wang C, Amin SA. Quorum sensing regulates ‘swim-or-stick’ lifestyle in the phycosphere. *Environ Microbiol* 2020; **22**: 4761–4778.
126. Comstock J, Henderson LC, Close HG, Liu S, Vergin K, Worden AZ, et al. Marine particle size-fractionation indicates organic matter is processed by differing microbial communities on depth-specific particles. *ISME Communications* 2024; ycae090.
127. Windler M, Bova D, Kryvenda A, Straile D, Gruber A, Kroth PG. Influence of bacteria on cell size development and morphology of cultivated diatoms. *Phycological Res* 2014; **62**: 269–281.
128. Bolch CJS, Subramanian TA, Green DH. The toxic dinoflagellate *Gymnodinium catenatum* (Dinophyceae) requires marine bacteria for growth 1. *J Phycol* 2011; **47**: 1009–1022.
129. Mayali X, Azam F. Algicidal Bacteria in the Sea and their Impact on Algal Blooms. *Journal of Eukaryotic Microbiology* 2004; **51**: 139–144.
130. Coyne KJ, Wang Y, Johnson G. Algicidal bacteria: a review of current knowledge and applications to control harmful algal blooms. *Front Microbiol* 2022; **13**: 871177.
131. Paul C, Pohnert G. Interactions of the algicidal bacterium *Kordia algicida* with diatoms: regulated protease excretion for specific algal lysis. *PLoS One* 2011; **6**: e21032.
132. Van Tol HM, Amin SA, Armbrust EV. Ubiquitous marine bacterium inhibits diatom cell division. *ISME J* 2017; **11**: 31–42.
133. Sison-Mangus MP, Jiang S, Tran KN, Kudela RM. Host-specific adaptation governs the interaction of the marine diatom, *Pseudo-nitzschia* and their microbiota. *ISME J* 2014; **8**: 63–76.



## Chapter 3

### **Marine protists as the base of evolution of animal-associated symbionts and pathogens**

---

Fabian Wittmers<sup>1,2</sup>, Camille Poirier<sup>2</sup>, Olga Matantseva<sup>3</sup>, Rex Malstrom<sup>6</sup>, Craig A. Carlson<sup>5</sup>,  
Lisa Sudek<sup>1</sup>, Stephen J. Giovannoni<sup>4</sup>, Alexandra Z. Worden<sup>1,2,3</sup>

<sup>1</sup>Marine Biological Laboratory, Woods Hole, MA 02543, USA

<sup>2</sup>Ocean EcoSystems Biology Unit, RD3, GEOMAR Helmholtz Centre for Ocean Research, 24148 Kiel, DE

<sup>3</sup>Max Planck Institute for Evolutionary Biology, 24306 Plön, DE

<sup>4</sup>Oregon State University, 97331 Corvallis, OR, US

<sup>5</sup>University of California Santa Barbara, 93106 Santa Barbara, CA, US

<sup>6</sup>DOE Joint Genome Institute, 94720 Berkeley, CA, US

## Abstract

Predatory protists occupy key positions in marine food webs with major implications for the flow of energy and carbon in open ocean ecosystems. Their interactions with other microbes beyond classical predator-prey relationships are poorly known, although much of eukaryotic evolution revolves around entanglements between unicellular phagotrophs and prokaryotes. Here, using single cell sorting in the wild, we reveal three lineages of uncultivated symbionts associated with uncultivated predators in the ocean, and identify a host for the recently named Gamibacteraceae (Rickettsiales, Alphaproteobacteria), a family of putative symbionts that lacked any known host association. The gammaproteobacterial Marinicoxiellaceae and Protistiellales as well as the alphaproteobacterial Simplirickettsiaceae and Gamibacteraceae are found throughout low and mid-latitude oceans and reside within choanoflagellates and MAST lineages, dominant consumers of the most abundant phytoplankton and other microbes. The symbionts association with choanoflagellates, the closest unicellular relatives of animals and members of the Opisthokonta which houses all metazoans, together with their evolutionary position based on phylogenomic analyses shed light on the emergence of four different groups of metazoan pathogens, including causative agents of disease in human, and specialized adaptations in the host environment. All four choanoflagellate and MAST-3 associated lineages possess complete Type IV secretion systems, however their manner of symbioses differ, following distinct trajectories. The Protistiellales show signs of horizontal transfer of genes from distantly related bacterial pathogens, in what appear to be preferential association with hosts, similar to *Legionella pneumoniae*. The Simplirickettsiaceae showed extensive reduction to a genome size <1 MB, with a loss of key parts of metabolism, including lysine biosynthesis, ribosomal proteins, and tRNA modification enzymes, suggesting an obligate symbiosis with their choanoflagellate hosts. Our results emphasize the need to examine symbiont and pathogen evolution in the context of optimal, permissive, and accidental hosts, and shifts that occur between these relationships. The widespread nature of symbioses in marine predatory protists expose complexities in resource flow in the oceans and point to possibilities for host jumping to multicellular animals, marine and otherwise. The diversity and evolutionary position of the novel symbiont-host relationships of marine protists uncovered could enable a better understanding of the evolution of bacterial pathogens of metazoans, as well as the ecology and complex resource flow observed within complex ocean foodwebs.

## Introduction

The oceans are known for the dominance of microbial life in controlling biogeochemical cycles critical to the global biosphere [1]. Predatory protists are a key link in these cycles and ocean food chains, consuming both phytoplankton and prokaryotes. Bacteria (many of them uncultured) are involved in a variety of processes, responsible for the respiration of organic matter within the microbial loop [2], but also include the photosynthetic cyanobacteria and diverse symbionts of protists and multicellular organisms alike [1]. Over evolutionary time, physical interactions between such phagotrophic protists and prokaryotes have had profound influences on eukaryotic life as we know it, including endosymbiosis events and the establishment of organelles. Indeed, protist-bacterial symbioses underpinned the rise of the algal progenitor of land plants and its capacity for photosynthesis [3, 4] and a plethora of mutualistic symbioses exist between modern day photosynthetic protists and bacteria [5, 6].

The impact of symbiotic bacteria on their eukaryotic hosts ranges from neutral to mutualistic or pathogenic [7], oftentimes varying depending on the host encountered. In metazoans, mutualistic symbionts of ticks provide essential nutrition to complement a blood diet, but cause disease upon entering mammalian hosts [8]. Manifestation of tick-borne bacterial diseases have so many similarities that they were thought to come from single type of bacterium, until the discovery that these pathogens actually represent evolutionarily distant lineages, specifically Alphaproteobacteria and Gammaproteobacteria [9–11]. Microbiomes of metazoans have now been extensively characterized and encapsulate interactions of profound consequence, leading to health or even death [12].

In contrast, while the diversity and feeding activities of key surface ocean phagotrophic protists have long been examined, their microbiomes are unknown and difficult to conceptualize given the small cell sizes of some taxa [13]. Detection of extant symbioses between marine predatory protists and bacteria has been hindered by the fragility of many protists, preventing easy recovery of intact cells and co-associates, as well as recalcitrance to culturing, the elimination of particular interactions upon culturing, and the inability of most molecular methodologies, such as metagenomics, to capture intact interactions [1]. An aquaria-based study recently exposed symbioses between two alphaproteobacterial lineages and marine diplomonads [14], protists that can switch between osmotrophy and phagotrophy [15]. Additionally, the one study to examine the microbiome of active predatory protists in the field, exposed an obligate symbiosis between an uncultivated Gammaproteobacterium (Comchoanobacterales) and choanoflagellates, widespread predatory protists that are the closest living unicellular relative of animals [3]. There are multiple other bacterial lineages,

characterized only by environmental metagenome assembled genomes (MAGs), that are proposed to have host-dependent lifestyles due to reduced genomes, but for which a true connection to hosts is not known. Prominent examples are the candidate phylum Omnitrophota, ultrasmall bacteria ( $\sim 0.2 \mu\text{m}$ ) based on size fractionated data, that are found in soils and aquatic environments, and the alphaproteobacterial family Gamibacteraceae [16, 17]. Both have complete central metabolism, but reduced genomes and lack a known host. This hinders our understanding of their impact on and the fate of their hosts (should they indeed be symbionts), their role in the microbial food web, and subsequent impact on the carbon cycle.

Symbioses open new ecological opportunities that have shaped the diversification of life [18–20] and continue to reveal new aspects of ocean biogeochemical cycling [5]. Bridging knowledge gaps on biogeochemical processes and potential future transitions depends on gaining more nuanced understanding of marine microbiomes. For protists, this requires extending our understanding of their ecology beyond predator-prey dynamics, to other interactions, such as the impact of viruses and co-associations with bacteria. It also requires disentangling roles of the many diverse uncultivated ocean bacteria for which lifestyles remain unknown.

## Material & Methods

Analysis in this chapter is based on Alphaproteobacteria and Gammaproteobacteria individual protist-associated metagenome assembled genomes (ipaMAGs) recovered in association with heterotrophic protists. For details on sampling, sequencing, and genome recovery, refer to the materials & methods section in chapter 2.

### Completeness and contamination refinement of high quality ipaMAG-0934

CheckM predicted two missing marker genes (PF01702, PF01259) and the duplication of two marker genes (PF00691, PF00152) for ipaMAG-0934, leading to a completeness estimate of 99.05% and 0.71% predicted contamination. Manual follow-up analysis confirmed that one of the duplicates of PF00152 (tRNA synthetases class II) was a horizontal gene transfer event of a transposase into a long contig (466 Kb), not representing contamination. PF00691 (OmpA) represents an intra-genome duplication event of a gene also found duplicated in other ipaMAGs described here (see below), encoded within two large contigs (466 Kb and 219 Kb), both of which are clearly part of the ipaMAG-0934 assembly. The absence of both PF01702 (Queuine tRNA-ribosyltransferase) and PF01259 (SAICAR synthetase, phosphoribosylaminoimidazole-succinocarboxamide synthase) is a feature of the novel family represented by ipaMAG -0934 (see below, Figure 8). The absence of this marker gene is also seen in the other members of this family discussed herein but is generally well conserved in Rickettsiales otherwise. This confirms that it has been lost within members of this lineage specifically and is not an indication of an incomplete genome assembly. This leads us to assume that ipaMAG-0934, based on marker gene analysis is a complete genome assembly that contains no contamination.

### 16S rRNA gene phylogenies

Specialized reconstructions were performed for the Rickettsiales and classical Legionellales. For the classical Legionellales, taxon sampling started based on a previously published phylogeny [21]. The sequence selection was updated by blasting the 16S rRNA gene sequences of ipaMAG-0156 and ipaMAG-0636, two high quality ipaMAGs classified as Gammaproteobacteria related to the Coxiellales based on GTDB, against NCBI nr. Sequences were considered for the alignment when having a minimum length of 1400 bp and being in the top 25 hits of at least one of the query sequences. Additionally, GTDB was queried for novel clades identified within this gammaproteobacterial lineage, and 16S rRNA

gene sequences were recovered from those, as well as all other species representatives of already included taxa using checkM 'ssu-finder' as described above. When encoding multiple 16S rRNA gene sequences, the longest sequence was considered for the alignment. Details on sequences included in the phylogeny as well as metadata and corresponding references are available as supplementary table 4A.

For Rickettsiales, initial taxon sampling was based on prior publications [17, 22]. We then screened all Rickettsiales assemblies classified as representative species in GTDB for 16S rRNA gene sequences as described above. Additional sequences were recovered and incorporated in the alignment by blasting 16S rRNA gene sequences for ipaMAG-0934 and ipaMAG-0929, two high quality ipaMAGs classified as Rickettsiales based on GTDB, against NCBI, as described above.

For both phylogenies, selected 16S rRNA gene sequences were aligned using the MAFFT (v7.310; [23]) L-INS-i algorithm [24] to avoid over-alignment and subsequently trimmed using trimAl (v1.4.1; [25]). Trimmed alignments were further manually inspected for anomalies and curated, removing chimeric and incomplete sequences. Phylogenetic trees for both lineages were reconstructed using IQ-TREE (v.1.5.4; [26]) with ultra-fast bootstrap (-b 1000) and optimized substitution models (Rickettsiales: GTR+F+I+R5; Gammaproteobacteria: GTR+F+R9) based on ModelFinder's [27] bayesian information criterion. Information on host-lineage and environment displayed in figures is based on NCBI metadata obtained from the nucleotide archive for 16S rRNA clonal sequences and NCBI BioSample data for 16S rRNA gene sequences recovered from assemblies. Phylogenetic trees and accessory data were visualized using ggplot2 (v3.4.2; [28]) with the ggtree extension (v3.6.1; [29]).

#### Phylogenetic reconstruction of horizontal gene transfer through gene trees

ZupT taxon sampling is based on sequences assigned to the protein family model for the zinc transporter ZupT (NF003243.0) as of 2 Oct 2023, in addition to the top 100 blastp hits for the ipaMAG-0636 ZupT protein sequence against NCBI nr (E-value <  $1 \times 10^{-10}$ ). Bax Inhibitor-1/YccA family protein (Bax1) taxon sampling is based on the NCBI protein family NF024221.0, as well as the top 100 blastp hits recovered using the ipaMAG-0636 Bax1 protein sequences as a query against NCBI nr (E-value <  $1 \times 10^{-10}$ ). Following the same approach, aminoglycoside phosphotransferase family protein (APH\_6\_hur) reconstruction was based on NF016533.0 and uridine diphosphate-N-acetylglucosamine-binding protein

YvcK on TIGR01826.1 as well as the respective blastp hits for ipaMAG-0156 query proteins. Sequences from NCBI protein family models were clustered at 80% sequence identity using CD-HIT (v4.8.1; [30]) with ‘-aL 0.80 -aS 0.5’ settings before combining with protein blast results. Sequences were aligned using MAFFT with default settings followed by trimming using trimAl with ‘-gt 0.3’. The phylogenetic reconstruction was performed using FastTree (v2.1.11; [31]), with ‘-lg -cat 3 -gamma’ settings based on 2452 sequences with 271 positions (ZupT), 2870 sequences and 296 positions (Bax1), 5184 sequences and 255 positions (APH\_6\_hur), and 3383 sequences and 307 positions (YvcA) respectively. Taxonomy was derived from NCBI using taxonomy ID and taxon to lineage information (downloaded 10 Feb 2023).

### Phylogenomics

Phylogenomic reconstructions of Gammaproteobacteria and Rickettsiales are based on orthologs datasets build for Legionellales (taxID: 118969) and Rickettsiales (taxID: 766) by eggNOG (v5.0; [32]). For Rickettsiales, initial taxon sample was based on [17, 22], with other genomes included to capture six additional families, specifically, Arcanobacteraceae, UBA6177, 33-17, JADKJE01, CAJLX01 and JABSSG01 (renamed herein Simplirickettsiaceae fam. nov.) as well as ipaMAG-0934, ipaMAG-0047 and ipaMAG-0929 recovered herein. Gammaproteobacteria taxon sampling was based on Needham et al. 2022 [21] with the addition of representatives from each genus of Coxiellales classified by GTDB, and all high-quality representatives of the uncultured M30B25 (named herein Protistiellales ord. nov.), PIVX01, UBA1113, UBA9339, UBA6002, UBA12402, and CAISUE01 lineages, and ipaMAG-0156, ipaMAG-0789 and ipaMAG-0636 generated herein. Orthologs were identified using hmmscan (E-value < 1 x 10<sup>-20</sup>) against the prebuilt eggNOG hmm database for the respective lineage to identify proteins with single-copy orthologs (SCOs) in ≥94% of the protein catalogs from the two custom genome collections, resulting in 90 gammaproteobacterial SCOs and 88 Rickettsiales SCOs that were considered further. Highly discordant SCOs were removed by performing single protein reconstructions using IQ-TREE (v2.2.6; options: -B 1000 -bnni -m MFP) and alignments generated in MAFFT E-INS-i (v7.520; --auto; [23]) that had been trimmed via trimAl with the -gappyout option (v1.4.rev15; [25]). For each resulting tree bipartition counts were generated using ‘tre\_make\_splits.pl’ using IQ-TREE bootstrapping output in scripts obtained from ‘github.com/novigit/broCode’, followed by an all-vs-all pairwise computation

of discordance using ‘tre\_discordance\_two.pl’. The 10% most discordant SCO based on average discordance against all other SCOs were removed resulting in 81 (classical Legionellales and related lineages) and 79 (Rickettsiales) SCOs used for subsequent phylogenomic reconstructions. In some cases, this approach led to removal of desirable taxa, for example *Ca. Arcanobacter lacustris* (GCA\_000970895.1), represented by a SAG that was 41% complete and lacked 36 of the 79 markers. This left a single representative of the Arcanobacteraceae, a MAG (GCA\_000746585.2) with all 79 markers.

For the lineage specific phylogenomic reconstructions based on the 81 and 79 SCOs identified herein, non-homologous sites were masked for each SCO by pre-alignment of sequences using prequal (v1.02; [33, 34]), aligned using MAFFT E-INS-i. SCO alignments were then trimmed and ‘divvied’ using Divvier (v1.01; [34]) removing columns with more than 60% gaps and additionally filtering with ‘-partial’. Separate SCO alignments were concatenated using ‘catfasta2phym.pl –concatenate’ (github.com/nylander/catfasta2phym). The untreated concatenated alignment was further trimmed to remove compositionally heterogeneous sites using ‘alignment\_pruner.pl --chi2\_prune f0.01’, to filter out the top 1% most heterogeneous sites iteratively as established in [35] with 67 (Rickettsiales) and 65 (classical Legionellales) repetitions respectively, until no taxa were deviating significantly in a  $\chi^2$ -test conducted after each iteration. Maximum likelihood phylogenies were reconstructed using IQ-TREE with 1000 ultra-fast bootstrap iterations under the Q.yeast+R6 (Rickettsiales) and LG+I+R7 (classical Legionellales) model as selected by ModelFinder [27] in IQ-TREE. Within the Rickettsiales, the Gamibacteraceae freshwater clade containing 42 MAGs was represented in the alignment by two members and was expanded in the visualization to indicate the position of the whole clade. Because multiple orders within the classical Legionellales displayed signs of genome reduction, we investigated the lineage-specific loss of single copy marker genes within the Marinicoxiellaceae, Comchoanobacterales and Protistiellales, which would indicate underestimation of genome completeness by checkM. We identified the missing marker genes for each genome within the three lineages considered in phylogenomic reconstruction using checkM’s lineage workflow, as described above. This revealed 19 SCMGs missing in all Comchoanobacterales, 13 in Protistiellales, and 11 in the Marinicoxiellaceae, which were subsequently not considered in manually corrected estimates of genome completeness.

To investigate the effect of *Ca. A. lacustris* on the tree topology and statistical support of the branch structure, we constructed an additional alignment and phylogeny including *Ca. A. lacustris* with the same 79 SCO marker set, despite the large portion of missing positions

(Figure S1). Additionally, we created an alignment and phylogenetic tree considering only the 43 SCOs present in *Ca. A. lacustris* (Figure S2). To investigate potential long-branch attraction between ipaMAG-0156 and the MAG JAJRR01 (GCA\_024279495.1), we reconstructed two additional phylogenies using the same 81 SCO marker set, selectively removing one of the two members of the long branch within the Coxiellales (Figure S3).

### Ortholog analysis

Representative genomes from each lineage represented in phylogenomic reconstruction of classical Legionellales and Rickettsiales were manually curated for ortholog analysis, choosing the highest quality genome or MAG for each lineage. Orthogroups were generated using OrthoFinder (v2.5.5; [36]) with the ‘msa’ gene tree inference method utilizing ‘mafft’, ‘diamond’ as the sequence search program and ‘iqtree’ for tree inference. Presence-absence patterns detected through analysis of orthogroups was confirmed against all genomes considered in the phylogenomics through diamond blastp searches (E-value  $1 \times 10^{-30}$ ) using the OrthoFinder orthogroup sequences as queries.

### Data availability

Raw reads for single-cell sorts and the highest quality ipaMAG of each lineage are available via NCBI BioProject PRJNA1061664. Full and trimmed alignments for phylogenetic trees of Alphaproteobacteria and Gammaproteobacteria, as well as all recovered bins are available through FigShare (doi.org/10.6084/m9.figshare.25751283.v1). This Whole Genome Shotgun project has been deposited at DDBJ/ENA/GenBank under the accessions JBDOIT000000000, JBDOIU000000000, JBDOIV000000000, JBDOIW000000000. The versions described in this paper are version 1.

## Results & Discussion

### Discovery of globally distributed uncultivated marine bacterial lineages co-associated with predatory protists.

To gain insights into potential cell-to-cell interactions between key open ocean predatory protists and bacteria we captured individual cells of wild choanoflagellates and MAST-3. Both are consumers of bacteria and phytoplankton, feeding actively on *Prochlorococcus*, the most abundant photosynthetic organism in the oceans [37, 38]. Choanoflagellates belong to the Opisthokonta ‘supergroup’ which also includes metazoans, while MAST-3 belong to the Stramenopila. Using single-cell sorting and subsequent whole genome sequencing, we characterized microbes that were physically co-associated with these important predatory protists, circumventing culturing biases or disruption of *in situ* cell-to-cell interactions.

Four bacterial lineages emerged that were each represented by a high quality ipaMAG (90-100% complete, 0.0-0.6% contamination), two Alphaproteobacteria and two Gammaproteobacteria, with two additional ipaMAGs (ipaMAG-0047, ipaMAG-0789) representing close relatives of two high quality ipaMAGs, bringing the total of analysed ipaMAGs herein to six (Table 1). Their completion levels indicated multiple clones were present with each sorted protistan cell, given that publicly available assemblies from single bacterial cells (SAGs) are usually highly incomplete [21, 39]. The presence of multiple, potentially clonal cells associated with the sorted protistian cell likely led to genome redundancy that greatly increase the completeness of our assembly.

| genome ID   | clade                                      | protist associat. | compl. | cont. | size_Mb | est. G. size | # prot. | % GC |
|-------------|--------------------------------------------|-------------------|--------|-------|---------|--------------|---------|------|
| ipaMAG-0934 | Simplirickettsiaceae ( $\alpha$ -proteo.)* | <i>B.minor</i>    | 100    | 0.0   | 0.97    | 0.97         | 937     | 39.1 |
| ipaMAG-0929 | Gamibacteraceae ( $\alpha$ -proteo.)       | <i>B.minor</i>    | 91.5   | 0.0   | 1.43    | 1.56         | 1322    | 36.9 |
| ipaMAG-0047 | Gamibacteraceae ( $\alpha$ -proteo.)       | <i>B.minor</i>    | 84.9   | 2.9   | 1.65    | 1.88         | 1478    | 36.0 |
| ipaMAG-0156 | Marinicoxiellaceae ( $\gamma$ -proteo.)*   | MAST-3            | 94.6   | 0.6   | 1.70    | 1.93         | 1484    | 44.7 |
| ipaMAG-0636 | Protistiellales ( $\gamma$ -proteo.)*      | <i>B.minor</i>    | 90.3   | 0.3   | 1.40    | 1.60         | 1341    | 41.4 |
| ipaMAG-0789 | Protistiellales ( $\gamma$ -proteo.)*      | <i>B.minor</i>    | 71.2   | 5.3   | 1.12    | 1.51         | 1131    | 42.3 |

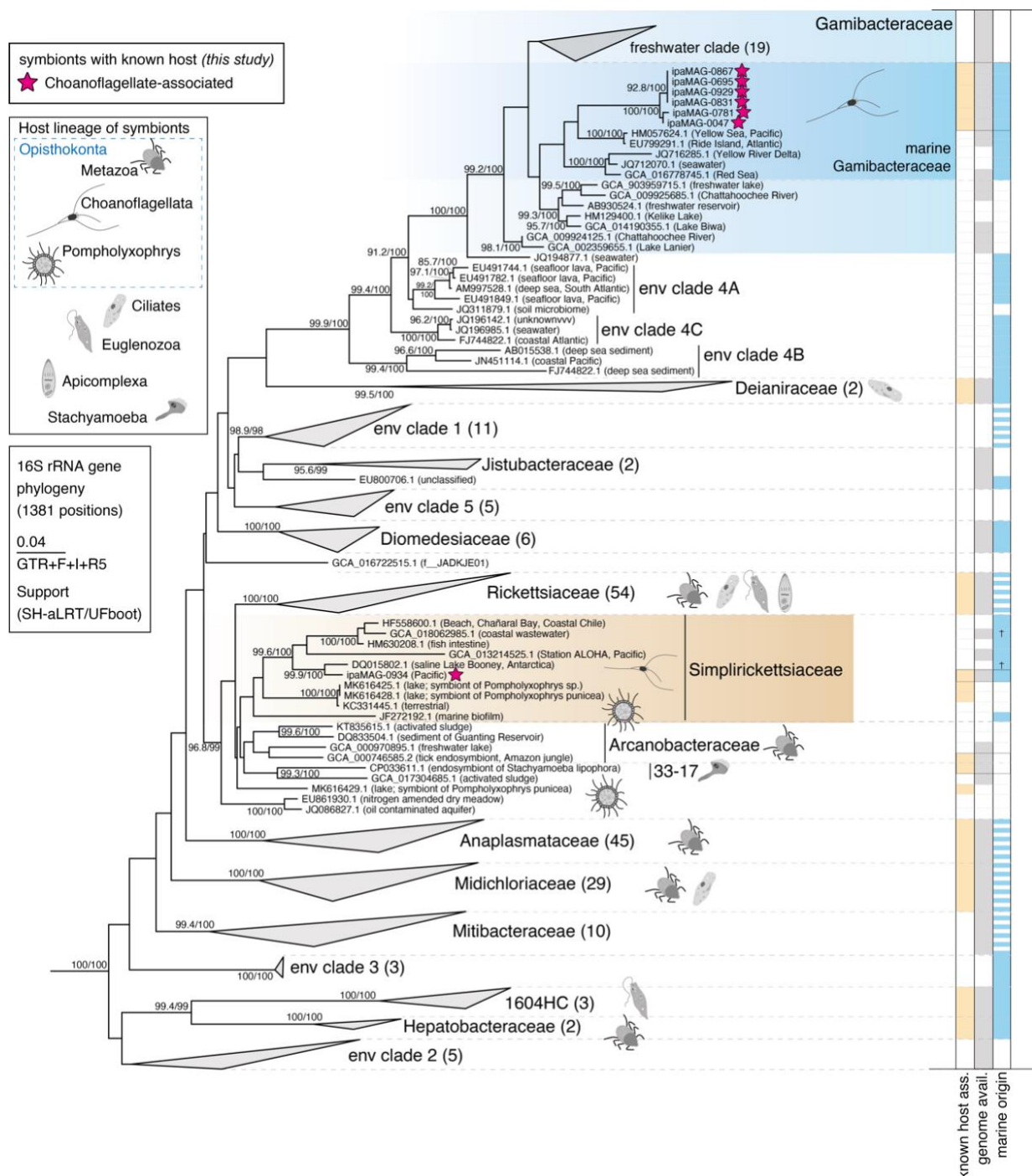
**Table 1: Assembly statistics for ipaMAGs from Alphaproteobacteria and Gammaproteobacteria, recovered from *B.minor* (Choanoflagellata) or MAST-3 SAGs. Clades marked (\*) represent uncharacterized lineages that have been named herein.**

Phylogenetic analysis of ipaMAG 16S rRNA genes established that they are divergent from all cultivated bacterial species. Three fell within two distinct clades comprised exclusively of marine environmental sequences in the alphaproteobacterial Rickettsiales

(Figure 1). ipaMAG-0929 and ipaMAG-0047 both branched together with marine representatives of the Gamibacteraceae, together with additional 16S rRNA sequences from lower-quality assemblies recovered from additional choanoflagellate single-cell sorts. ipaMA-0934 formed a deep branch forming a sister-clade to the Rickettsiaceae. The others fell within what has traditionally been termed the (gammaproteobacterial) Legionellales (ipaMAG-0636, ipaMAG-0156, Figure 2). A fifth group recovered in co-association with choanoflagellates belonged to the recently described Comchoanobacterales [21] (Figure 2) and was not further analysed herein. We also recovered two full-length 16S rRNA gene sequences (*B. minor* sort\_0716 & sort\_0016) related to ipaMAG-0636 (ipaMAG-0789 did not contain a 16S rRNA), suggesting that the diversity of Choanoflagellate symbionts from that lineage goes beyond the ipaMAG described herein, but we were unable to recover a high quality ipaMAG for sort\_0716. All four of the novel protist-associated groups had genomes sizes <2 Mb and tended towards AT-richness (low %GC, Table 1), compositional features typical of eukaryotic endosymbiotic bacteria [40]. IpaMAG-0934 was just 966 Kb, comparable to well characterized Rickettsiales endosymbionts with among the smallest bacterial genomes known, such as *Wolbachia* (864 Kb) [41] and *Neorickettsia* (859 Kb) [42]. These features pointed to possible symbiosis between the uncultivated bacterial lineages having symbiotic relationships with the unicellular predators sorted in the wild.

#### Reflections of novel marine bacterial-predatory protist co-associates on the evolution of animal pathogens.

Understanding of the diversity, evolution, and function of human pathogens like those within the Rickettsiales (Figure 1) and classical Legionellales (Figure 2) is largely based on strains from animals, including insects (e.g.,[43]), or protistan model systems, like the ciliate *Paramecium* [44]. Comparisons often center around related pathogens like *Coxiella* or *Legionella* [45, 46], with a few notable recent exceptions [14, 17, 22, 47]. Closely related environmental clades are typically excluded, many of which are not represented by high quality genome assembly or lack the contextual information (such as whether they are host associated) that is crucial for understanding co-evolution with eukaryotes.



**Figure 1: Phylogenetic relationship of Rickettsiales highlights environmental clades with unknown lifestyles.** Taxonomic sampling is based on lineages characterized in Castelli [22] and Schön [17]. Additional sequences were recovered based on blastp searches of ipaMAG 16S-rRNA gene sequences against NCBI nr and by screening all genome of species representative in GTDB ‘o\_Rickettsiales’ for complete 16S rRNA gene sequences. ML reconstruction based on full-length (including all sequences from cell sorts) and near full-length (minimum 1400 bp) 16S rRNA gene sequences. Trimmed alignment with 1381 positions was evaluated based on GTR+F+I+R5 model. Node support is based on Shimodaira-Hasegawa approximate likelihood-ratio test (SH-aLRT; [48]) and 1000 ultra-fast bootstrap (UFboot) replicates and support is indicated for clades with SH-aLRT  $\geq 80\%$  and UFboot  $\geq 95\%$ . The tree was rooted using five *Pelagibacter* (SAR11) 16S rRNA gene sequences (not displayed in visualization).



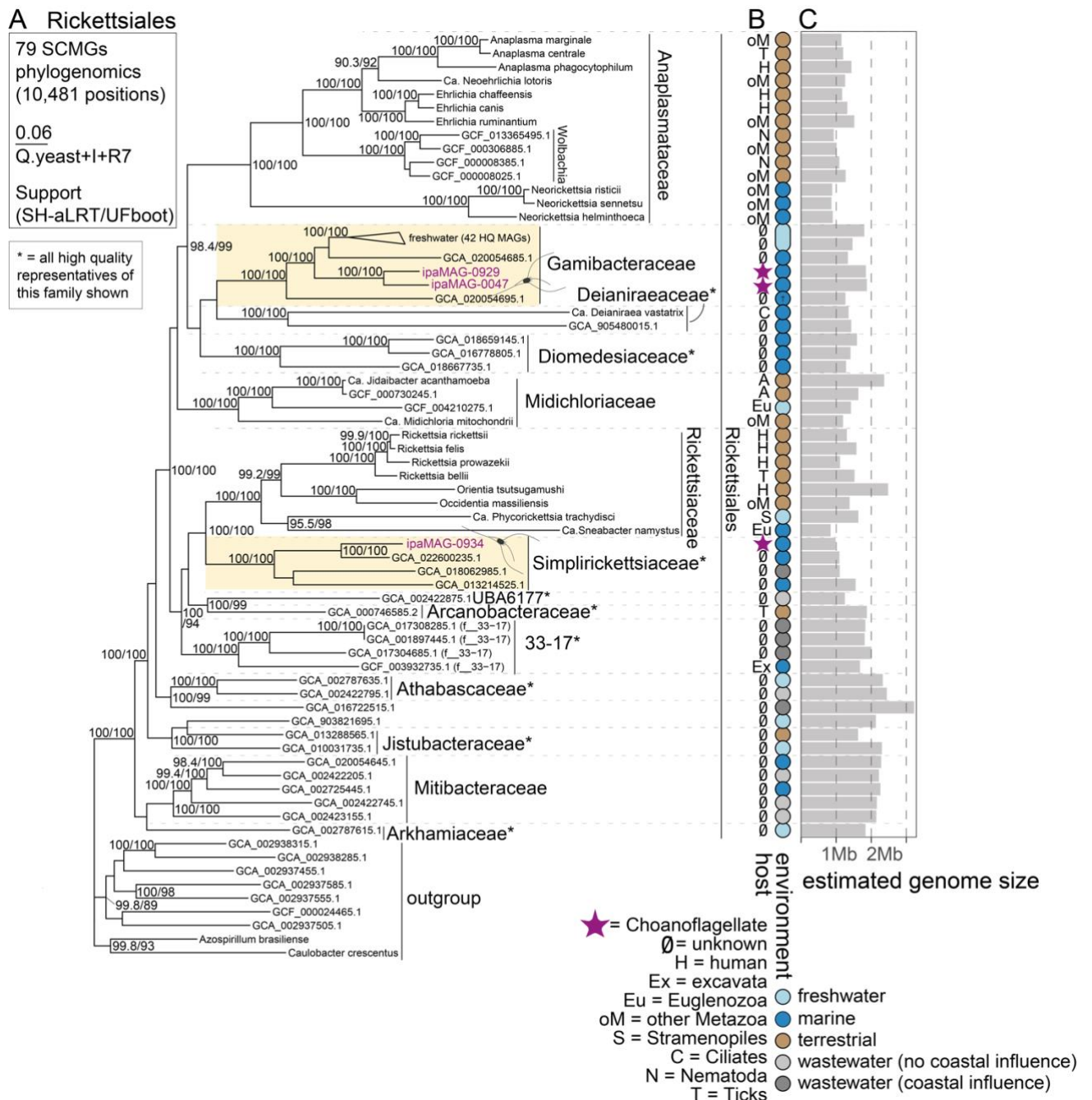
**Figure 2: Phylogeny of Coxiellales and related uncultivated and symbiotic gammaproteobacterial orders.**

Taxonomic sampling is based on lineages in [21]. Additional sequences were recovered based on blastp searches of ipaMAG 16S-rRNA gene sequences against NCBI nr and by screening all genome of species representative in GTDB related to Legionellales and Coxiellales for complete 16S rRNA gene sequences. ML reconstruction based on full-length and near full-length (minimum 1400 bp) 16S RNA gene sequences. The trimmed alignment with 1486 positions was evaluated using the GTR+F+R9 model. Analysis of the 16S rRNA gene from the Marinicoxiellaceae genomes placed them in the same clade, embedded within a diverse supported clade of other marine environmental sequences, including UBA9148, all from uncultivated organisms. Node support is based -on SH-aLRT [48] and ultra-fast bootstrap replicates and support is indicated for clades with SH-aLRT  $\geq 80\%$  and UFboot  $\geq 95\%$ . Accessions for all sequences, including those in collapsed clades, as well as metadata and references are available in supplementary table 4. The tree was rooted using members of the Francisellales as an outgroup (omitted in the visualization).

We next constructed phylogenomic trees, often used when characterizing deep-branching lineages and allowing to include MAGs that often lack full-length 16S rRNA gene sequences and can subsequently not be considered for 16S rRNA gene-based phylogenies. To infer genome scale phylogenies of the Rickettsiales and classical Legionellales that would also reveal the evolutionary position of the new host-associated lineages, we performed comprehensive taxonomic sampling with two aims. First, while attempting to incorporate taxonomic representation of all relevant lineages in GTDB [49], we also ensured that environmental MAGs from other studies were close to completion and had low contamination, factors that influence the phylogenetic relationships recovered, statistical support, and evolutionary distances. Second, taxonomic sampling of available genomic data was agnostic to environmental or medical importance.

Phylogenomic reconstruction of the Rickettsiales showed one choanoflagellate-associated lineage is the sister clade to the Rickettsiaceae, based on analysis of 79 SCMGs conserved across the Rickettsiales (Figure 3). The core Rickettsiaceae include *Rickettsia rickettsia* and *Rickettsia prowazekii*, causative agents of Tick-borne Rocky Mountain Spotted Fever and Louse-borne Typhus, respectively, alongside other pathogens, as well as symbionts of non-Opisthokonta protists [50, 51]. We named the novel lineage the Simplirickettsiaceae, because the ipaMAG-0934 genome was smaller than many symbionts and pathogens in the Rickettsiaceae ( $1.4 \pm 0.3$  Mb; Figure 3), apart from the divergent symbiont of marine diplomonids, *Ca. Sneabacter namystus* [14]. In the phylogenomic analysis Simplirickettsiaceae comprised uncultivated marine lineages, which until now lacked known

host-associations (Figure 3). In the 16S rRNA gene analysis we observed expansion to freshwater, including apparent symbiosis with a freshwater opisthokont (Figure 1).



**Figure 3: Phylogenomic tree of Rickettsiales.** Maximum-likelihood reconstruction of Rickettsiales based on 79 concatenated single copy ortholog alignments (10,481 positions) evaluated under the Q.yeast+I+R7 model. The two choanoflagellate co-associates (*ipaMAG-0929* and *ipaMAG-0047*) have 70% pairwise amino acid identity (AAI), higher than to other members of the Gamibacteraceae. Node support is based -on SH-aLRT [48] and ultra-fast bootstrap replicates and support is indicated for clades with SH-aLRT  $\geq 80\%$  and UFboot  $\geq 95\%$ . Alphaproteobacterial *ipaMAGs* are highlighted in violet. Accessory data include (B) information on host-association and environment, as well as (C) estimated genome size.

The environmental 33-17 family holds an evolutionarily important position in this region of the Rickettsiales tree. It comprises coastal wastewater MAGs with unknown lifestyles and a symbiont of *Stachyamoeba lipophora* ATCC 50324 [47], a predatory protist isolated from the Atlantic Ocean that belongs to the Excavata. Notably, support for topology was challenged by the Arcanobacteraceae and UBA6177 which lack a sufficient number of high-quality MAGs (Figure 3). The position of these families adjacent to the core Rickettsiaceae and Simplirickettsiaceae was not supported, however the Arcanobacteraceae were previously reported to be the closest sister-clade of the Rickettsiaceae [52]. Additional reconstructions examining these instabilities confirmed the statistically supported sister relationship between the Rickettsiaceae and Simplirickettsiaceae (Figure S1, S2). With the caveat that many sequences in this region of the tree come from uncultivated bacteria that lack any lifestyle information, our findings suggest symbioses with aquatic protists strongly influence and inform on key aspects of the Rickettsiaceae ancestor. Moreover, ipaMAG-0934 established that the Simplirickettsiaceae are host-associated, specifically symbionts of opisthokonts, like many of the Rickettsiaceae, and the complete genome offered new possibilities for understanding both the evolution of the choanoflagellate associated lineage but also the *Rickettsia* themselves.

The second choanoflagellate-associated alphaproteobacterial lineage formed a statistically supported branch within the recently discovered Gamibacteraceae [17]. Genomic features of this aquatic lineage point to symbiosis with eukaryotes, although lacking definitive host information [17]. Our data establish choanoflagellates as a host and insights that expand upon the limited information for marine members. Phylogenomic analysis recovered a freshwater Gamibacteraceae clade, and four marine members, two ipaMAGs and two assembled from waters collected offshore from South Africa. Support for backbone topology in this region of the Rickettsiales tree was lacking at several nodes but was present between the Anaplasmataceae, symbionts of insects, including *Wolbachia*, and pathogens of animals, and a group gathering the Gamibacteraceae, Deianiraceae (which has been implicated in long branch attraction issues [17]), and another uncultivated marine family, the Diomedesiaceae [22]. While the latter lack known host-associations, *Ca. Deianiraea vastatrix* was isolated from a marine ciliate [53]. Our expanded taxonomic sampling, restricted to high-quality MAGs, indicates that the “crown” Rickettsiales partitioned to two broad clades, one harbouring the Anaplasmataceae and the above aquatic clades and the other harbouring the Rickettsiaceae, our Simplirickettsiaceae, as well as other environmental clades like the

[illegible]

71

**Figure 4: Phylogenomic trees of classical Legionellales.** (A) Maximum-likelihood based phylogenomic reconstruction of 79 gammaproteobacterial assemblies based on 81 single copy ortholog alignments concatenated resulting a single alignment of 12,607 positions, reconstructed under the LG+I+R7 model. Gammaproteobacterial ipaMAGs are highlighted in violet. Note that JAJRRL01 and ipaMAG-0156 share 85% of their protein encoding genes. Node support is based -on SH-aLRT [48] and ultra-fast bootstrap replicates and support is indicated for clades with SH-aLRT  $\geq 80\%$  and UFboot  $\geq 95\%$ . Accessory data include (B) information on host-association and environment, as well as (C) assembly size (dark grey) and estimated missing portion of the genome (light grey) for each genome.

The gammaproteobacterial ipaMAGs alongside other environmental MAGs comprised independent lineages within the classical Legionellales, which encompasses animal pathogens such as *Legionella pneumophila* and *Coxiella burnetii*, the former also residing in aquatic protists (*Acanthamoeba*), and the latter a tick-symbiont (Figure 4). MAST-3 co-associate ipaMAG-0156 and JAJRRL01, an Indian Ocean deep-sea hydrothermal vent MAG, formed a sister clade to the core *Coxiella*, including *C. burnetii*. We term these the Marinicoxiellaceae, acknowledging that the deep branching structure between ipaMAG-0156 and JAJRRL01, the sole representative of the GTDB order of the same name, will lead to further discovery and delineation. In the 16S rRNA gene analysis, a bacterial symbiont lacking genomic information was the closest sequence to ipaMAG-0156 (Figure 2). The symbiont was isolated from a marine sponge [54], which like choanoflagellates and metazoans, belongs to Opisthokonta [55]. These results indicating that closely related members of the Marinicoxiellaceae can be symbionts of Opisthokonts and Stramenopiles. Given the few HQ MAGs available for the Marinicoxiellaceae, we also investigated whether long branch attraction might influence this topology. We performed two additional phylogenomic analyses (Figure S3) and examined AAI for the entire Legionellales lineage (Figure S4), which supported relatedness of these two deep branching Marinicoxiellaceae. The reconstruction placed multiple other aquatic environmental lineages, many marine and all lacking lifestyle information, in statistically supported branching positions adjacent or basal to the *Coxiella* and Marinicoxiellaceae.

The choanoflagellate-associated gammaproteobacterial ipaMAG-0636 branched in a different region of the tree than the Marinicoxiellaceae. It formed a supported clade named here the Protistiellales, that included a second ipaMAG from a wild choanoflagellate (ipaMAG-0789) and the MAG M30B25 (lacking a 16S rRNA gene sequence) from a saline lake [56], environments known to have abundant choanoflagellates [57]. The Protistiellales grouped with the order PIWD01 (represented by a single marine MAG) and uncultivated

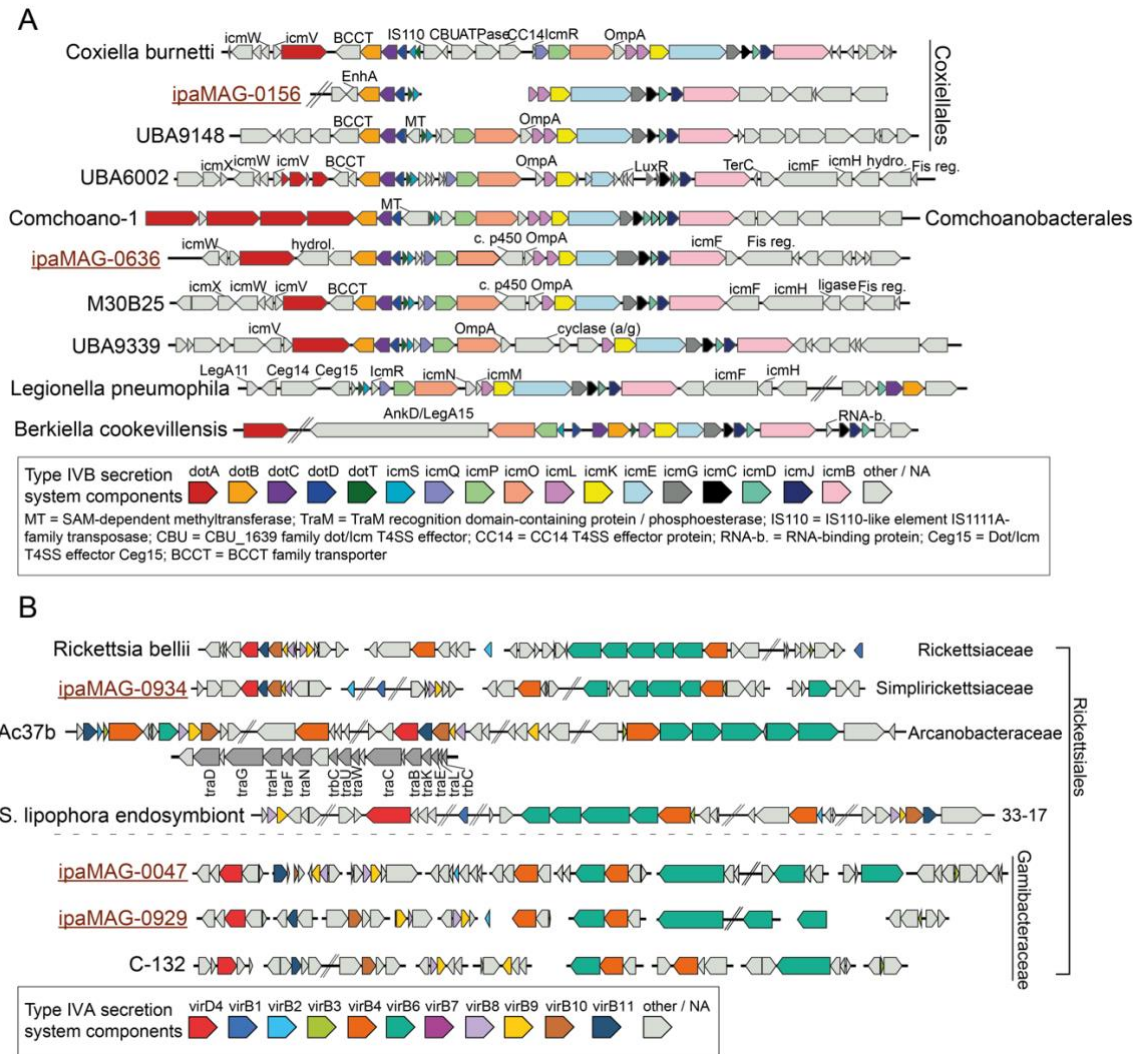
obligate symbionts of choanoflagellates, Comchoanobacterales [21]. Protistiellales formed a supported sister group to the latter. However, its estimated genome size ( $1.52 \pm 0.07$  Mb), indicates Comchoanobacterales' extensive genome reduction (1.01 Mb) [21], represents a unique evolutionary trajectory towards complete host dependency, and manifests differently than the Protistiellales, Marinicoxiellaceae as well as the many other aquatic lineages that comprise the massive diversity covered by the classical Legionellales.

The overall background topology of this portion of gammaproteobacterial diversity still requires resolution that will be improved as more lineages are discovered. Here, groups like the Diplorickettsiales are in unsupported positions basal to the diversity encapsulated by the *Coxiella*, Marinicoxiellaceae, Protistiellales and multiple other principally aquatic lineages. These are collectively more related to each other than they are to the 'true' Legionellales (Figure 2, 4). Supported regions of our topology contrast sharply with analyses that do not include the newly discovered Marinicoxiellaceae and Protistiellales, or exclude other lineages recovered from the aquatic environments. Indeed, key nodes recovered and used to determine a pathogenic and animal macrophage infecting ancestor of the breadth of Legionellales tick endosymbionts appears highly improbable. This would require that said ancestor also spread to marine protistan hosts, and then re-emerged to infect terrestrial animals in multiple independent events. Our analysis point to multiple events where symbionts of aquatic protists have led to mutualistic symbioses in animals, such as insects. These same symbionts then cause disease upon encountering other hosts, consistent with the hypotheses put forward by Duran et al. [58].

It is well known that several bacterial lineages harbouring animal pathogens have members that are symbionts of aquatic protists, often protists with diverse microbiomes [12, 59], and all belonging to 'supergroups' other than the Opisthokonta. *L. pneumophila* itself is found in *Acanthamoeba* and *Paramecium*, and individual members of the Rickettsiaceae have been isolated from a Euglenozoan (diplonemid) and a stramenopile alga [14, 50]. To our knowledge the *Coxiella* and the Anaplasmataceae have only been reported in association with animals (including insects). Our finding of multiple independent uncultivated bacterial lineages that live in association with marine protists, especially Opisthokonta, provides a new perspective for ease of transition from one opisthokont host to others, potentially involving transit through permissive hosts from other eukaryotic supergroups. Our identification of the Marinicoxiellaceae ipaMAG-0156 extend the host range of *Coxiella*-like organisms from primarily metazoan-associations to unicellular stramenopiles (MAST-3).

#### Nuances in mechanisms and metabolic approaches implicated in symbiosis.

A key feature of pathogenesis and symbiosis alike is the Type IV Secretion system (T4SS) [60]. Of the two T4SS superfamilies, one is essential for injecting bacterial effector proteins into the host, and the other's function is connected to conjugation [60]. We observed T4SS of the former in all four uncultivated bacterial lineages (Figure 5). The Simplicirickettsiaceae had highest synteny and amino acid identity to Rickettsiaceae, consistent with the phylogenomic results and vertical inheritance. Likewise, the two Pacific Gamibacteraceae were more similar to each other than to the best prior marine Gamibacteraceae MAG (C-132) and had higher amino acid identities to evolutionarily closer orders (Deianiraeaceae, Anaplasmataceae and Midichloriaceae) than to Rickettsiaceae or Simplicirickettsiaceae. All four novel lineages perform glycolysis, with variations connected to the Entner-Doudoroff glycolysis pathway, lacking phosphofructokinase, whereas the Rickettsiaceae and Comchoanobacterales lack glycolysis completely (Figure S5). None of the four novel predator-associated lineages are flagellated, although generally, many common marine Alphaproteobacteria and Gammaproteobacteria are motile [61, 62].



**Figure 5: Synteny of type IV secretion systems including ipaMAGs, symbionts and environmental lineages.** T4SS of the classical Legionellales (A) and Rickettsiales (B) and lineages identified herein, both of which are in the Effector Translocators Superfamily. Note gene size scaling is different between the two panels; gaps in MAG assemblies are denoted by a gap in the horizontal line between contigs. Selected contigs were collapsed (//) for visualization purposes. Note that members of the Arcanobacteraceae family (here represented by Ac37b) also encode a Conjugation Superfamily T4SS, not found in any other lineage shown herein, and depicted below the Effector Translocator T4SS.

Simplirickettsiaceae is the most distinct from the other three symbiont lineages found herein, which share more similarities. Simplirickettsiaceae encodes an ATP/ADP translocase, a classical sign of energy parasitism that enables uptake of ATP from the host [63]. This translocase is also present in the Rickettsiaceae and Comchoanobacterales [21] and other obligate intracellular symbionts [64]. Of the four lineages, only Simplirickettsiaceae lacks a competence system, which is completely or partially present in all Rickettsiaceae, and partially in Comchoanobacterales (ComEC). Three of the lineages can biosynthesize essential

B-vitamins that function as co-enzymes, although none encoded B12-related pathways. The choanoflagellate-associated Gamibacteraceae encode *de novo* B1 biosynthesis and salvage pathways, as well as full pathways for B2, B5, and B7, similar to many Anaplasmataceae. Marinicoxiellaceae biosynthesize B2, B5, and B7, but not B1. *Coxiella* species are similar, although a few synthesize B1, and the B5 pathway appears to be partial or absent [65]. The Protistiellales also synthesizes B2, B5, and B7, and can condense the two thiamin moieties to form TPP, the biologically active form of vitamin B1 (thiamin) [66] but cannot produce the precursors. Vitamin B1 is scarce in the ocean photic zone, despite being an essential for all domains of life. Several key lineages of marine bacteria and phytoplankton cannot fully biosynthesize it, nor can choanoflagellates (or animals) [67, 68]. Simplicirickettsiaceae ipaMAG-0934 has reduced B2, and completely lacks B5, B7 and B1 biosynthesis (like Rickettsiaceae), and thus is nearly as reduced as Comchoanobacterales which lack biosynthesis of all B-vitamins. Vitamin exchange has been implicated as a vital part of symbiont-host mutualism in multiple systems, including provision of B-vitamins by *Wolbachia* to blood-feeding (insect) hosts [69]. In this sense MAST-3 and choanoflagellates may use symbioses to solve a shared problem – vitamin supplies.

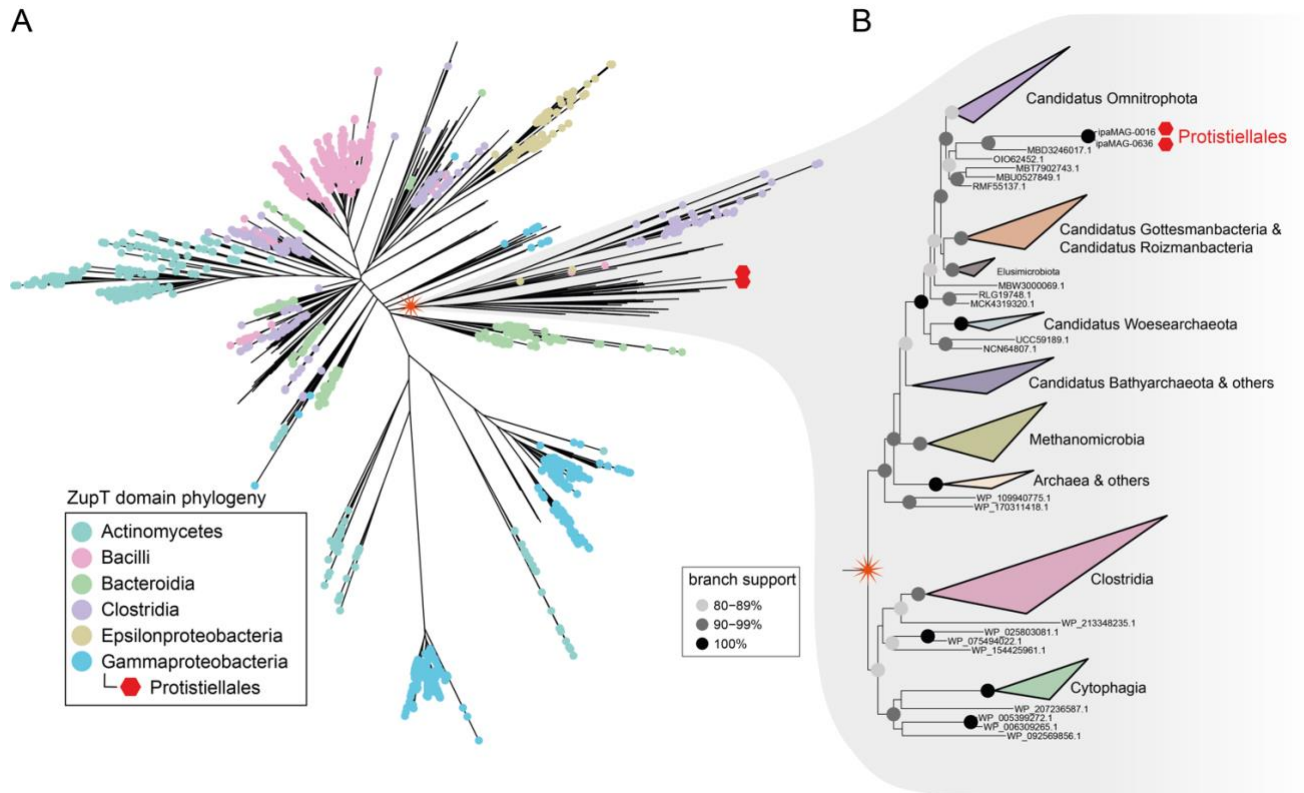
#### Trajectories of symbioses with marine predatory protists – independent routes.

Convergent evolution of symbionts from different bacterial lineages in the same or similar hosts, including *Acanthamoeba* [70], diplomonads [14], and ants [71] has been observed before. Our results support this at a wider level, with examples such as vitamin biosynthesis but at the same time highlight how individual ‘unique’ HGT events have led to specialized aspects of the host-symbiont interaction, that differ across the disparate lineages. Given that three of the novel lineages were co-associated with the same hosts, and the fourth (Marinicoxiellaceae) seems to infect opisthokonts as well, but also stramenopiles, this appears in some way due to niche specialization or potential mechanism that evade competition with other symbionts or well-developed host immune responses. Here, several symbiont gains-of-function link to cell surface properties and direct interaction with hosts – but, although largely operating in shared hosts, are functionally different.

Each of the predatory protist-affiliated lineages harboured novel approaches to host interaction. The two Gamibacteraceae recovered in association with choanoflagellates, encode a 5-histidylcysteine sulfoxide synthase OvoA, which catalyses formation of 5-histidylcysteine sulfoxide, a small sulfur compound involved in signalling and oxidative

stress responses [72]. It is implicated in defence against host immune responses for intracellular eukaryotes, specifically trypanosome parasites and symbionts (members of the Euglenozoa) [72, 73]. While not detected by orthology-based searches in other Gamibacteraceae or most other Rickettsiales (possibly due to incomplete MAGs) it was also present in the Diomedesiaceae and in multiple other bacterial phyla (as well as eukaryotes, including choanoflagellates, some Metazoa and some unicellular stramenopiles). In bacteria it is disproportionately present in symbionts (including parasites and pathogens) of animals and plants, where it is implicated in metabolic exchanges between ovothiol-producing bacteria and hosts lacking this function [74].

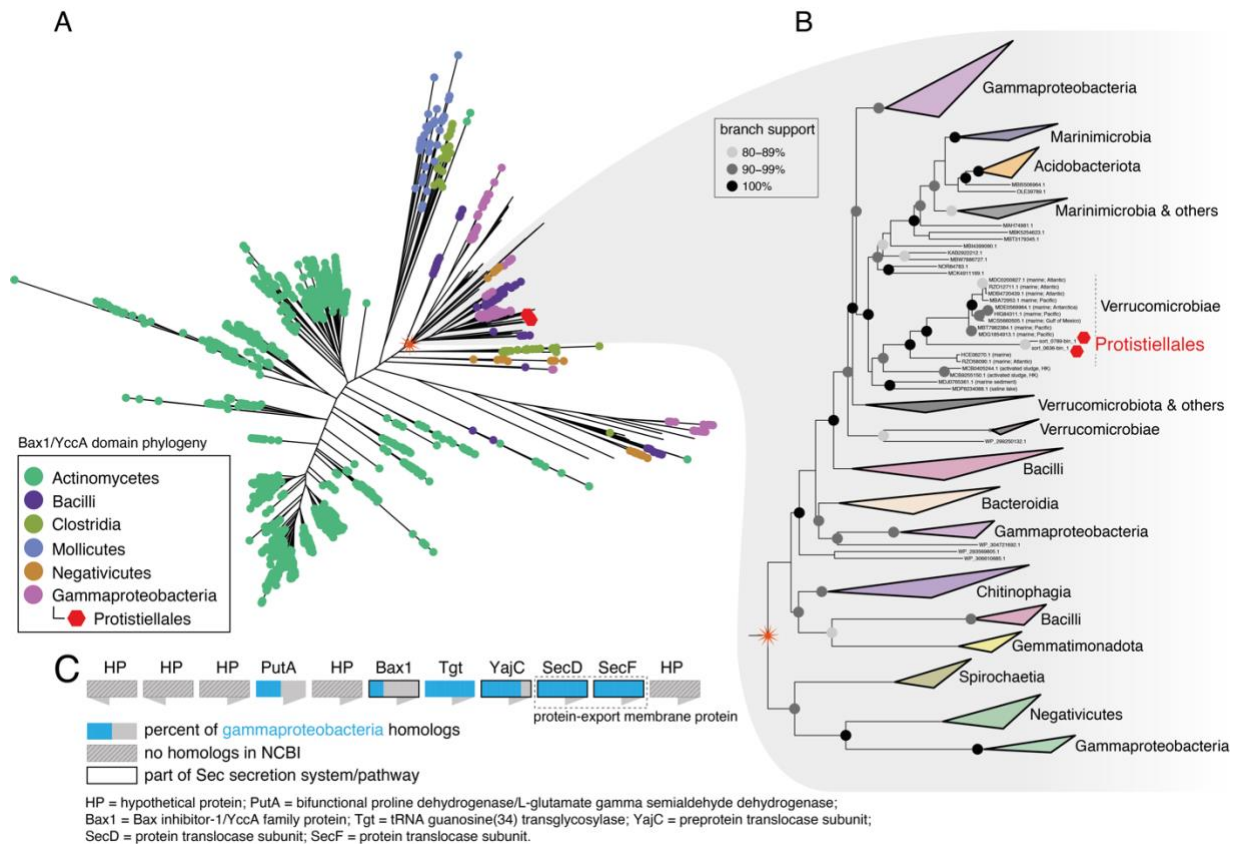
Protistiellales ipaMAGs encode a series of proteins not found in any related lineage of endosymbiotic pathogens from the classical Legionellales. This includes multiple proteins associated with virulence of pathogens, such as a zinc transporter ZupT [75] as well as a Bax inhibitor-1/YccA family protein [76, 77]. ZupT mutant work in cultures demonstrates that it helps *Clostridium difficile* counteract host-imposed nutritional restriction (through transitional metal sequestration) [78]. It also has a role in success of fish pathogens from multiple bacterial lineages, including the Gammaproteobacterium *Aeromonas salmonicida* [75]. Here, phylogenetic reconstruction indicates ZupT was acquired through HGT (Figure 6). The Protistiellales ipaMAG (but not the saline lake Protistiellales MAG M30B25) proteins are in a statistically supported clade formed otherwise of ZupT from uncultivated bacterial phylum Omnitrophota and Woese archaeota, as a sister to a large clade comprising Omnitrophota ZupT. These ultra-small (<0.2µm) bacteria are found across diverse environments, including marine ecosystems and genomic profiling suggests the majority have a host-associated lifestyle [16]. ZupT acquisition by Protistiellales demonstrates a possible example of antagonistic coevolution of Protistiellales with its choanoflagellate host, or simply a means by which zinc requirements of both host and symbiont can be met.



**Figure 6: Horizontal acquisition of ZupT.** (A) Phylogenetic maximum-likelihood-based reconstruction of the ZupT domain. Taxon sampling is based on blastp homologs of the Protistiellales ZupT sequences and sequences from the NCBI protein model domain NF003243.0. Taxonomy information is based on NCBI taxon IDs and node support (based on UFboost) is not depicted in the global version (A) of the tree for visualization purposes. (B) represents the subsection of the phylogenetic tree that contained the Protistiellales ZupT sequences.

Protistiellales encodes three OmpA proteins (the bacterial outer membrane protein A), used by *Coxiella burnetii* during host cell invasion [79]. While OmpA is present throughout the classical Legionellales, including Marinicoxiellaceae, the three Protistiellales OmpA appear to have been acquired horizontally from other sources, contrary to the vertically inherited OmpA present throughout the related lineages. One is most similar to OmpA in the gammaproteobacterial *Acinetobacter*, a human pathogen found in soil and water environments [80], while the other two are located adjacently in the genome and show homology to various Alphaproteobacteria. The fact that OmpA differs from those in nearby relatives with symbiont/pathogen lifestyles likely reflects host range expansion or differentiation. While OmpA has been implicated in multiple roles, the Protistiellales also acquired a Bax1/YccA-family protein with closest homologs being uncultured marine Verrucomicrobiota (Figure 7). Implicated as a virulence factor for infection and triggering of host immune response by *Edwardsiella piscicida*, a gammaproteobacterial fish pathogen [81, 82], here Bax1/YccA is directly upstream from inner membrane proteins of the Sec pathway

(protein export) and just upstream are four hypothetical proteins with no functional predictions and no homologs in genbank nr.



**Figure 7: YccA phylogeny.** (A) Phylogenetic maximum-likelihood-based reconstruction of the ZupT domain. Taxon sampling is based on blastp homologs of the Protistiellales Bax1 sequences and sequences from the NCBI protein model domain NF024221.0. Taxonomy information is based on NCBI taxon IDs and node support (based on UFboost) is not depicted in the global version (A) of the tree for visualization purposes. (B) represents the subsection of the phylogenetic tree that contained the Protistiellales Bax1 sequences. (C) Organisation of the ipaMAG-0636 (Protistiellales) contig encoding Bax1. Also indicated are results using the Protistiellales sequences as a queries, reflecting the proportion of sequences (blue shading) in the top 10 blastp results that were assigned to a gammaproteobacterial assembly.

Protistiellales, as well as marine group UBA9148 encode a homolog of the proteorhodopsin found in Comchoanobacterales, with the same core motif. Microbial rhodopsins have been noted for their high abundance, and for a few bacterial taxa have now been demonstrated to allow phototrophy [83–85]. While in Comchoanobacterales their function is still unclear, possible roles include energy transfer and acidification of the host cell [86]. Like Comchoanobacterales, homology-based searches indicate the Protistiellales is lacking the pathway for beta-carotene biosynthesis, the pigment that is cleaved for production of the necessary retinal chromophore [21]. This indicates the chromophore must be recovered

from choanoflagellate prey cells – or potentially choanoflagellate giant viruses that encode both rhodopsin and the pigment required [87]. The similarity of the protein in these gammaproteobacterial lineages indicates that the last common ancestor of the Coxiellales (given proteorhodopsin presence in UBA9148), Comchoanobacterales and Protistiellales encoded a proteorhodopsin that was lost in classical *Coxiella* as well as Marinicoxiellaceae, and multiple other nonmarine lineages. The evolutionary origin of this horizontally acquired protein could not be reconstructed due to the short sequence length and high mutation rate, i.e., divergence, observed within rhodopsin evolution.

ipaMAG-0156 (Marinicoxiellaceae) carries two putative Rpn (Recombination-Promotion Nuclease)-like transposase proteins [88]. The Rpn system is a toxin-antitoxin module encoded in bacteria, in which one protein encodes the toxin, and the other suppresses toxin activity, until a trigger causes release and toxin activation. This function was recently demonstrated in *E. coli* (replacing prior proposed functions as some form of transposable element endonuclease) [89]. Phage infection triggers activation, in what appears to be an abortive infection strategy whereby the cell is either sacrificed or rendered dormant, in the interest of preventing infection of the greater population [89]. Rpn systems are present in various bacterial taxa, including many Rickettsiales, but is not in Simplicirickettsiaceae or Gamibacteraceae and has not been reported in Coxiellales until now. Ortholog analyses show that while not present in other *Coxiella*, it is present in scattered representatives in more basal lineages, specifically a Legionellales, a Diplorickettsiales, a UBA1113 and a PIVX01 and in GCA\_018402195.1 (the saline lake MAG from the Protistiellales). However, the ipaMAG-0156 Rpn system has more similarity to those encoded in *Citrobacter* species and *Escherichia coli*, than more closely related Gammaproteobacteria like Legionellales, suggesting a horizontal acquisition. Both Rpn systems detected here (RpnA and RpnB) contain the characteristic four amino acid repeats comprising an oligomerization domain within the C-terminal region, which encodes a separately transcribed protein (Rpn<sub>S</sub>) within the full protein (Rpn<sub>L</sub>). Interestingly, Rpn<sub>B</sub><sub>L</sub> does not encode the four amino acid repeats, but an almost exact copy of the C-terminus of Rpn<sub>B</sub><sub>L</sub> is encoded downstream of Rpn<sub>B</sub><sub>L</sub>, containing an insertion of 10 four amino acid repeats, suggesting that the Rpn<sub>S</sub> component of this system is encoded on a separate open reading frame. The repeats in Rpn<sub>B</sub><sub>S</sub> also deviated in their normal composition of one hydrophobic and three charged amino acids or glycine, seen in RpnA and all characterized *E.coli* systems, being composed of QQGM instead. It has been suggested that Rpn<sub>L</sub> is involved itself in its duplication, as well as the creating of smaller downstream copies and amino acid repeats [89], but this and the exact role of this putative

phage defence system, especially within symbiosis between bacteria and eukaryotic host organisms remains to be investigated.

IpaMAG-0156 and its deep branching sister MAG (JAJRR01, GCA\_024279495.1) within the Marinicoxiellaceae uniquely share two proteins connected with interactions and cell wall structure, relative to other Coxiellales. One is an aminoglycoside phosphotransferase (APH\_6\_hur) characterized as being involved in antibiotic resistance [90], and the other is uridine diphosphate-N-acetylglucosamine-binding protein YvcK, involved in cell wall integrity [91]. Each protein appears to have been acquired by the LCA of these two lineages, but in different events, although limited taxonomic sampling of Marinicoxiellaceae (the only two representatives being ipaMAG-0156 and JAJRR01) prevent conclusive determination (Figure S6). Apart from these acquisitions and the Rpn system, ipaMAG-0156 does not show extensive acquisition through HGT or loss of genes, consistent with the suggestion that the influences of HGT were most substantial in the *Coxiella* ancestor, prior to divergence the subclades within the lineage [43, 46]. In fact, 93% of ipaMAG-0156 orthogroups also contained *Coxiella* members, suggesting a large overlap in encoded genes overall.

Both Comchoanobacterales and the Simplirickettsiaceae seem largely recalcitrant to influx of new functionality that often occurs in parallel with genome reduction in symbionts. We did not observe clear acquisition of virulence factors, potentially due to loss of the competence system in Simplirickettsiaceae. Simplirickettsiaceae has retained complete pathways for lipid and lipopolysaccharide synthesis, synthesis of most of amino acids, and glycolysis, unlike Comchoanobacterales which has lost all of these [21]. However, Simplirickettsiaceae has lost various functions associated with resistance to host defence mechanisms and virulence factors, when compared to *Rickettsia* (Figure 8).

Our results establish much broader host-ranges for nearby relatives of many metazoan endosymbionts that lead to disease in some hosts. Likewise, their genomic content points to common solutions, as well as distinct approaches, for colonizing hosts. While the timing of acquisition events is difficult to ascertain, given extinction or absence of data from many lineages, it some of the HGT events observed herein could have occurred through the interactions between bacterial members of the protist microbiome, or through exposure to DNA from bacteria that are ingested by the predatory protist as prey. Exchanges of actual compounds between symbionts (and host) has also been clearly documented, for example in *Wolbachia* [92]. Possible co-residence of Simplirickettsiaceae with other bacteria occurs in the opisthokont *Pompholyxophyrs*, where Simplirickettsiaceae 16S rRNA genes have been seen [93]. An overlap, albeit in a small percentage of cells was observed for different

members of the Comchoanobacterales [21]. Such co-habitation could provide ongoing avenues for HGT from other prokaryotes, even if the symbiosis with the protist is obligate, although clearly less so for taxa that lack the competence system like Simplicirickettsiaceae. These interactions can also completely shift manifestation of the symbiosis or ‘infection’. For example, in the *Acanthamoeba* microbiome plays essential roles in consequences of infection. The host is lysed when infected by *Legionella pneumophila* alone, but when other bacterial endosymbionts are also present lysis is prevented [12]. Further, *Acanthamoeba* has been proposed as providing an environment where gene transfer between co-infecting symbionts occurs, influencing the development of effective pathogens of mammals, like the tick endosymbiont *Rickettsia bellii* [94]. It is likely that a simultaneous, intracellular infection of the same host organism provides close encounters that facilitate this transfer, presumably prior to loss of competence genes that has occurred in *Buchnera* [95] and some *Rickettsia* species [94]. Hence, there is much to learn about manifestation of infection when in a co-evolved host where mutualism may have been achieved, versus other hosts that are encountered where there are negative consequences to infection.

#### Gain and loss differentiate the Simplicirickettsiaceae and Rickettsiaceae

The recovery of ipaMAG-0934 provided unprecedented views into functions gained or retained in core Rickettsiaceae but absent in their sister, the Simplicirickettsiaceae. Despite its smaller genome size ipaMAG-0934 shares many similarities with Rickettsiaceae like *Rickettsia rickettsii*. However, 49 orthogroups present in the *Rickettsia* were absent from Simplicirickettsiaceae (with a specific focus on ipaMAG-0934) (Figure 8). Multiple cell-wall associated proteins and transporters (e.g., HlyC / CorC family MG<sup>2+</sup> exporter, EmrB/QacA family transporter, efflux RND transporter permease, autotransporter OmpB) were detected that have previously associated with intracellular pathogen invasion of hosts and resistance to host defences mechanisms [96–99]. Several of these were more broadly distributed across the Rickettsiales, like HlyD, and proteins with broad functions, but definitively connected to virulence in *Staphylococcus aureus* (CtS41 family peptidase) [100], another animal pathogen that proliferates in *Acanthamoeba*. The type 1 secretion system (T1SS) responsible for secretion of proteins and utilized in pathogenic interactions (HlyD secretion protein), were also present. Collectively, this set provides insight into the mechanisms acquired or retained by *Rickettsia* during their evolution towards symbiotic interactions with multicellular

organisms. It also highlights reductions in the Simplirickettsiaceae, such as what appears to be degradation of T1SS.

A striking example of *Rickettsia* differentiation from the broader Rickettsiales, and most dramatically so for its sister clade Simplirickettsiaceae, is synthesis of Queuosine precursors and presence of tRNA-guanine transglycosylase TGT, used in the final conversion to Queuosine (QueE, QueF, QueC, QueA, QueH). Queuosine forms a tRNA modification involved in diverse processes in animals, including cell proliferation, where it is acquired from gut bacteria and is used by the bacteria themselves [101]. ipaMAG-0934 and one other Simplirickettsiaceae lack all components, and the Anaplasmataceae and Midichloriaceae (largely tick endosymbionts) encode only TGT, reflecting a salvage-based strategy. Our results point to a differentiation in strategies across the symbiotic lineages of the Rickettsiales. The first involves complete biosynthesis in a strategy potentially linked either to coordinating expression of genes related to virulence and biofilm formation, as recently observed for several pathogenic bacteria [102], or an exchange maintaining benefits to a host having only blood for nutrition. The second strategy involves dependency on salvage, implying other nearby sources for both the host and symbiont, or less dependency on translational controls via this tRNA modification. The third type, complete loss, is seen in two of four Simplirickettsiaceae, including the choanoflagellate-associated symbiont, in *Sneabacter*, the marine diplomonid symbiont, and complete or partial loss in uncultivated marine member of the Deianiraceae. Complete biosynthesis (presumed retention from an ancestral free-living state) is a trait of *Rickettsia* species alone, opening new avenues for understanding *Rickettsia*-host interactions.

ipaMAG-0934 also encoded unique proteins not shared with closely related genomes. Perhaps as expected for a novel uncultivated symbiont lineage, most of the 78 proteins uniquely present in ipaMAG-0934 relative to the Rickettsiales recovered no functional annotations and indeed several appeared to be completely novel. Additionally, six unique orthogroups were observed in Simplirickettsiaceae compared to Rickettsiales, with only one having any functional annotation (a glycosyltransferase). Both essential enzymes of lysine metabolism are among losses that appear unique to Simplirickettsiaceae while among ipaMAG-0934's 78 unique proteins is a PQ-loop domain containing transporter which could transport lysine across the cell membrane [103].

The Simplirickettsiaceae ipaMAG-0934 is additionally lacking multiple proteins encoding core metabolic functions that are present in *Rickettsia*. These include TypA translational GTPase and ribosome biogenesis GTP-binding protein YihA/YsxC, both involved in

ribosome biogenesis, as well as proteins encoding core steps of purine biosynthesis, amino acid & phospholipid metabolism. ipaMAG-0934 is also missing ribosomal proteins S21 and L34, which are absent from some *Rickettsia* but not others. Ribosomal protein S21 has been categorized as non-ubiquitous and is frequently lost in bacteria [104]. It is however present throughout the Rickettsiales, apart from Simplicirickettsiaceae. Its dispensable nature is likely due to being near the ribosome surface and having lower number of contacts to other essential ribosomal components than many other ribosomal proteins. In contrast, while also surficial, L34 has only emerged as a candidate for loss through analysis of highly reduced bacterial genomes (<1 Mb, e.g., *Wolbachia* and *Neorickettsia*) [104]. Loss of L34 through mutation does impact growth in *B. subtilis*, in an offset that can be partially suppressed by magnesium [105], a cation second only to sodium in seawater. While these loss events are independent from those in the *Rickettsia*, some parallel reduction in the more distant Anaplasmataceae has been observed [41, 42]. Moreover, intense pressure on Simplicirickettsiaceae to reduce genome size is shared with evolutionary pressures in its sister clade, but in a manner differing from the gains and losses in *Rickettsia*.

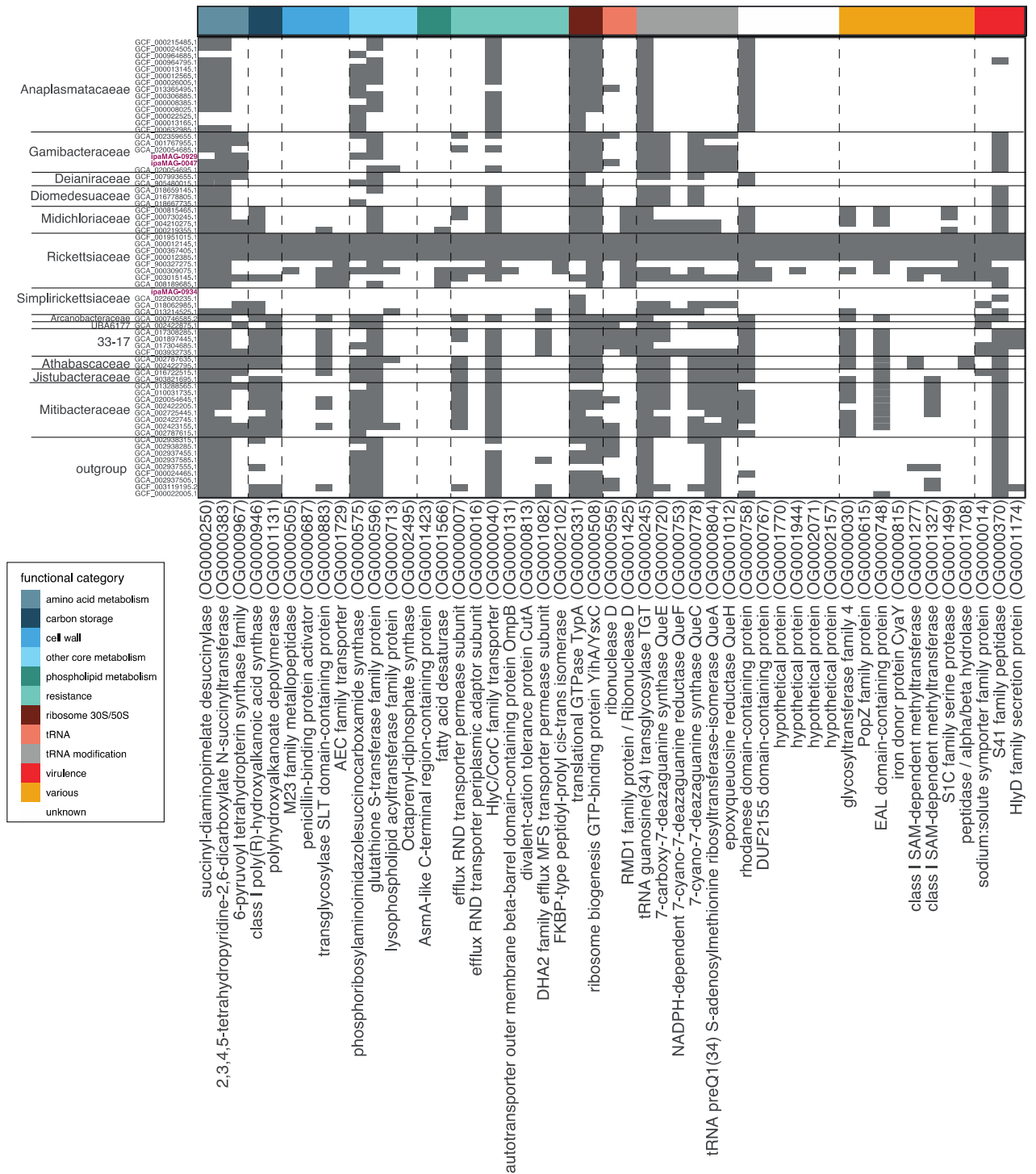
#### Environmental importance and the optimal host

Our results thus far demonstrated that some of the smallest protistan predators in today's oceans have symbionts, coming from divergent bacterial lineages. The types of symbioses represented by the four lineages living in association with these key predators with variations in the level of demand on the host. The Simplicirickettsiaceae, have the highest demand as obligate intracellular symbionts, while the Marinicoxiellaceae, Protistiellales, and Gamibacteraceae, are symbionts and seemingly bring more to their hosts, including synthesis of various B-vitamins, which are scarce in ocean surface waters [106]. The Gamibacteraceae have been proposed as obligate symbionts that operate either as ectosymbionts, consistent with lifestyle of the adjacent albeit deep branching Deianiraceae [53], ectosymbionts of *Paramecium*, or as intracellular symbionts [17]. A free-living or facultative symbiotic lifestyle is suggested for the more basal Mitibacteraceae and Athabascaceae [17], likely associated with protists, given estimated timing of the emergence of the Rickettsiales ancestor versus metazoans. Choanoflagellates themselves purportedly arose in the late Neoproterozoic [107], and while this timing does not align with proposals of an earlier Rickettsiales emergence, molecular dating is extremely challenging for prokaryotes and soft bodied eukaryotes. Interactions between bacteria and choanoflagellates have illuminated cell

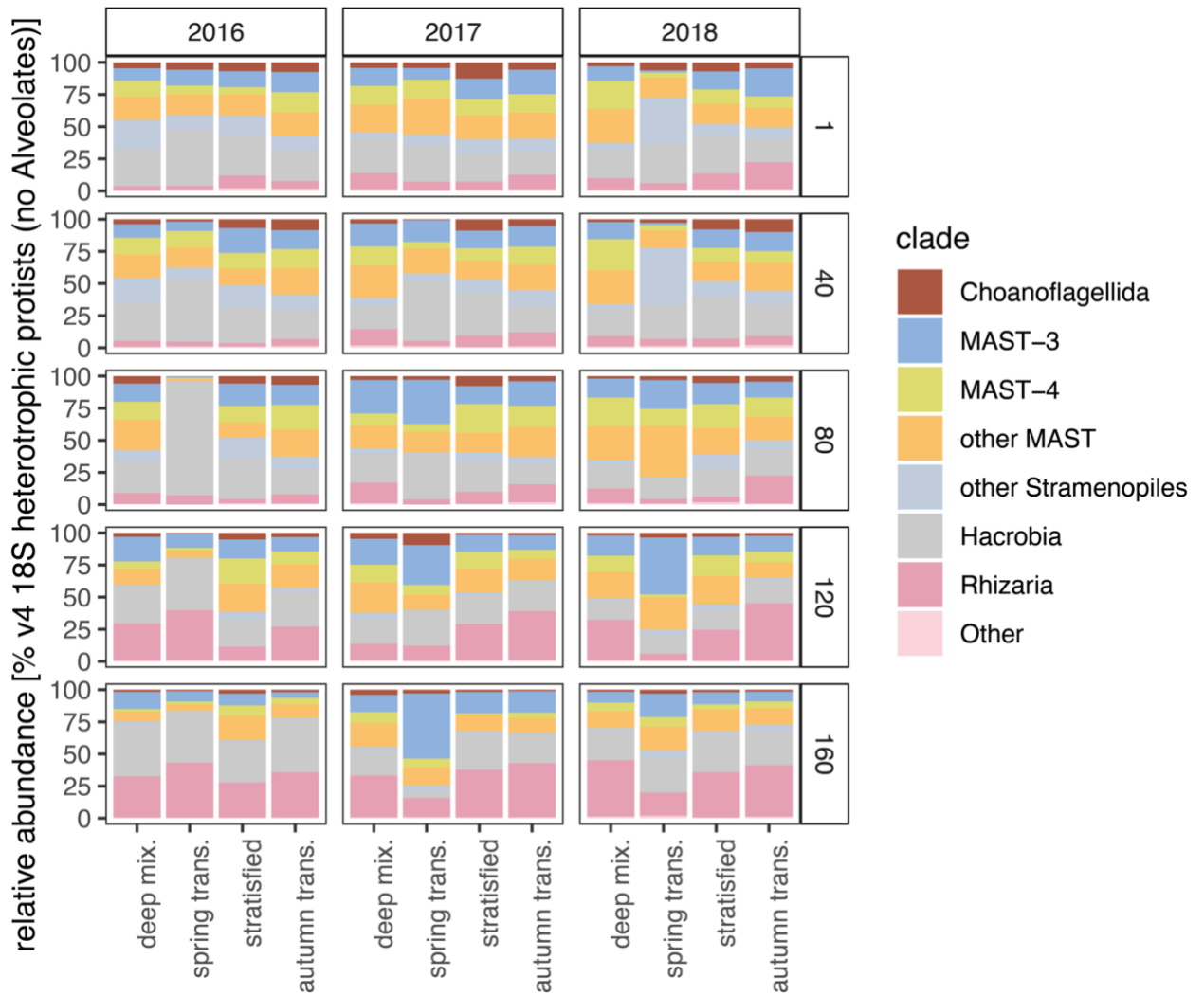
biological transitions central to animal evolution. Indeed, bacterial chemical queues trigger sex and the transition to and from a unicellular to multicellular state in choanoflagellates [108, 109]. The mechanism by which these queues arrive at high enough concentration if released in seawater remains mysterious and would be facilitated with contact-based exchanges. Our data point to the possibility of many more physical and symbiotic interaction between bacterial symbionts and small predatory protists, which herein are prominent for choanoflagellates.

The global distribution of all ipaMAGs from the four lineages suggests successful establishment of a sustained niche, much like *Coxiella*-infections of animals globally. The novel lineages were typically low in relative abundance (see chapter 1) compared to abundant free-living taxa, like *Prochlorococcus* and SAR11, making them statistically unlikely to be encountered at frequencies anything like common prey, and further raised questions about possible transmission and host prevalence.

Searches for the hosts at BATS, where corresponding 18S V4 amplicon data was available for a subset of samples, showed that choanoflagellates and MAST-3 were present throughout the photic zone and seasonal cycle (Figure 9), alongside previously reported broad distributions in global survey data [21]. MAST-3 are the most abundant MAST clade at the Bermuda Atlantic Timeseries Study (BATS) site and make up on average  $15.01 \pm 5.86\%$  of the non-photosynthetic protist community in the photic zone at across all seasons (recalculated using data from [110]). Moreover, the massive diversity shown for the marine symbiont lineages through 16S rRNA gene analysis (Figure 1, 2) suggests there are many more symbiotic relationships to be discovered. Collectively, the manner of recovery through capture of wild protist cells, characteristics of the high-quality ipaMAGs and the four lineages they represent, as well as their distributions throughout tropical and subtropical surface oceans, pointed to these bacteria having an important role in the ecology of dominant marine unicellular predators.



**Figure 8: Functions lost in ipaMAG-0934.** Presence-absence of orthologs that were found to be missing in ipaMAG-0934 (Simplirickettsiaceae) but present in core Rickettsiaceae. Lost functions were identified in orthofinder and presence-absence throughout Rickettsiales is based on diamond blastp searches (E-value < 1 x 10<sup>-30</sup> and 35% sequence identity) against all assemblies considered for phylogenomic reconstruction. Functions were manually group into categories, based on ghostkoala KEGG and diamond blastp annotations of the lost orthologs.



**Figure 9: Presence of Choanoflagellates and MAST-3 at the Bermuda Atlantic Time-Series station.**

Relative abundance of heterotrophic marine protist lineages at BATS based on V4 18S rRNA gene amplicon sequencing. Lineages were assigned to trophic modes based on Wilken et al. [38] and were normalized to the abundance of all heterotrophic protists excluding Alveolates. Relative abundances are averaged for each year and season, based on the BATS physical framework. Alveolates were excluded due to high numbers of Syndianales and the difficulties in assigning predatory versus mixotrophic modes to many dinoflagellates.

Many aspects of theory regarding symbiosis in the ocean with hosts that may be difficult to encounter suggest these are mutualistic relationships, or potentially highly flexible symbioses that entail multiple hosts, under current environmental conditions and ocean microbiome composition. We captured the associations by sorting active protistan predators in the surface ocean, however there are indications of expanded host ranges. For example, *Simplirickettsiaceae* 16S rRNA gene sequences were also detected in a freshwater unicellular opisthokont (Figure 1). For the *Gamibacteraceae*, in addition to having many freshwater members, 16S sequences were found in microbiomes from a multicellular opisthokont, the

cnidarian *Hydra vulgaris* [111], and a ciliate [112]. For the Marinicoxiellaceae, we detected 16S sequences in marine sediments and in a multicellular marine opisthokont (sponges). In contrast, the highly reduced obligate intracellular choanoflagellate symbiont Comchoanobacterales [21] appears restricted to hosts residing in marine pelagic environments. Its sister lineage in the phylogenomic reconstructions, Protistiellales, has much broader distributions. Indeed, close relatives of Protistiellales occur in deep sea sediments, soils, freshwater and even infected root canals (in human; Figure 1). Thus, a central question surrounds how the nature of these interactions may shift through climate change and encounter with other potential hosts.

Interestingly, until now, the Coxiellales have only been known to infect metazoans, a contrast to the Rickettsiaceae which are known symbionts of diverse eukaryotic lineages including metazoans, but also ciliates, Euglenozoa, Apicomplexa and others [113]. Given marine and other environmental subclades within both families that have little or no members with known host-association, it is likely that the host range of both is currently underestimated. Although there have recently been insights into the evolution within *Coxiella* as pathogens of humans and other Metazoans [46], the diversity within *Coxiella* as well as the larger phylogenetic position of the *Coxiella* clade within the Gammaproteobacteria is often oversimplified or neglected. This is driven largely by the absence of genomic and/or host information for multiple environmental lineages. We observe distinct surface proteins in Marinicoxiellaceae that might explain the associations with different hosts when compared to *Coxiella*. More generally, we observed extensive acquisition and loss of proteins associated to cell surface, the bacterial interface for symbiont-host interactions, which could partially explain the wide host ranges observed in the symbionts and pathogens of all the Rickettsiales and *Coxiella*-like lineages described herein (Fig. 4,5) [17]. Recent work has led to the discovery of unique members of common symbiotic lineages, such as *Sneabacter namystus* associated with diplonemids [14] and *Deianiraceae vastatrix*, an ectosymbiont of Paramecium [53]. Our findings add to these new lineages but are unique because they reveal symbionts of key lineages of widespread open ocean predatory protists.

Our results reveal symbionts of multiple marine nanoflagellates. Notably, these symbionts and their hosts do not associate in purely pairwise interactions. We found three instances in which we detected both Comchoanobacterales and Gamibacteraceae ipaMAGs within the same single choanoflagellate cell. It is likely that other lineages of symbionts also cooccur. Simplirickettsiaceae were not detected with another symbiont likely because the lineage is less common than Gamibacteraceae while Marinicoxiellaceae, associated with MAST-3 was

not detected more frequently because the dataset only comprised 9 MAST-3 cells. Symbiotic relationships can drastically change when another symbiont is associated with the same cell, as demonstrated in *Acanthamoeba* [12]. This suggests that bacterial interactions with heterotrophic protists should be considered conceptually similar to other microbiomes, with the symbionts themselves interacting, going beyond isolated interactions between a single symbiont and its host.

### Conclusions

Our currently understanding of the diversity of major lineages of human pathogens, classical Legionellales and Rickettsiales, is primarily shaped by symbionts of multicellular organisms, largely because they are comparatively easy to study and medically significance. This led to a skewed picture of the evolution and true diversity of these lineages in the larger scope, impacting our understanding of their evolution with subsequently biased interpretations as a consequence, harming progress in understanding these lineages better. Both lineages originated as symbionts of protists, highlighting the need to look at the microbiome of protists in general, including small heterotrophic nanoflagellates, not just larger protists, or multicellular eukaryotes, with important implications that go beyond microbial ecology. The true diversity of both lineages is likely far greater than currently known, with many (often protist-associated) lineages waiting to be characterized.

In the context of marine ecology, key predatory protists controlling populations of abundant phytoplankton like *Prochlorococcus* and like other bacterial groups alike, seem to harbour a microbiome, impacting the protists in ways currently poorly understood. Some show signs of beneficial functions such as vitamin biosynthesis, while some might be (energy) parasites. They likely affect growth both positively and negative (lineage-dependent) and flow of energy through these key components of the marine foodweb, highlighting the importance to investigate the symbiotic associations they are engaged in in more depth.

### **Author Contributions**

F.W. performed bioinformatics analysis, phylogenomics, phylogenetic reconstructions and all data visualisations. F.W., S.J.G., and A.Z.W. conceptualized the analysis. C.P. and A.Z.W. designed the sampling strategy and conducted single-cell sorting. L.S. and R.M. prepared single-cell sequencing plates and generated raw data. C.A.C. and O.M. provided feedback and conceptual ideas. F.W. and A.Z.W. wrote the manuscript.

## References

1. Worden AZ, Follows MJ, Giovannoni SJ, Wilken S, Zimmerman AE, Keeling PJ. Rethinking the marine carbon cycle: Factoring in the multifarious lifestyles of microbes. *Science* (1979) 2015; **347**: 1257594.
2. Azam F, Fenchel T, Field JG, Gray JS, Meyer-Reil LA, Thingstad F. The Ecological Role of Water-Column Microbes in the Sea. *Mar Ecol Prog Ser* 1983; **10**: 257–263.
3. Brockhurst MA, Cameron DD, Beckerman AP. Fitness trade-offs and the origins of endosymbiosis. *PLoS Biol* 2024; **22**: e3002580-.
4. de Vries J, Archibald JM. Plant evolution: landmarks on the path to terrestrial life. *New Phytologist* 2018; **217**: 1428–1434.
5. Tschitschko B, Esti M, Philippi M, Kidane AT, Littmann S, Kitzinger K, et al. Rhizobia–diatom symbiosis fixes missing nitrogen in the ocean. *Nature* 2024.
6. Foster RA, Tienken D, Littmann S, Whitehouse MJ, Kuypers MMM, White AE. The rate and fate of N<sub>2</sub> and C fixation by marine diatom-diazotroph symbioses. *ISME J* 2022; **16**: 477–487.
7. Husnik F, Tashyreva D, Boscaro V, George EE, Lukeš J, Keeling PJ. Bacterial and archaeal symbioses with protists. *Current Biology* 2021; **31**: R862–R877.
8. Biggs HM. Diagnosis and management of tickborne rickettsial diseases: Rocky Mountain spotted fever and other spotted fever group rickettsioses, ehrlichioses, and anaplasmosis—United States. *MMWR Recommendations and Reports* 2016; **65**.
9. Santos-Garcia D, Morel O, Henri H, El Filali A, Buysse M, Noël V, et al. Genomic changes during the evolution of the Coxiella genus along the parasitism-mutualism continuum. *Peer Community Journal* 2023; **3**.
10. Massung RF, Dasch GA, Ereemeeva ME. Chapter 17 - Rickettsia and Coxiella. In: Budowle B, Schutzer SE, Breeze RG, Keim PS, Morse SA (eds). *Microbial Forensics (Second Edition)*. 2011. Academic Press, San Diego, pp 277–295.
11. Raoult D, Marrie TJ, Mege JL. Natural history and pathophysiology of Q fever. *Lancet Infect Dis* 2005; **5**: 219–226.
12. Lena K, Cecilia W, Frederik S, Florian W, Sarah E, S SM, et al. Symbiont-Mediated Defense against Legionella pneumophila in Amoebae. *mBio* 2019; **10**: 10.1128/mbio.00333-19.
13. Zubkov M V, Tarran GA. High bacterivory by the smallest phytoplankton in the North Atlantic Ocean. *Nature* 2008; **455**: 224–226.

14. George EE, Husnik F, Tashyreva D, Prokopchuk G, Horák A, Kwong WK, et al. Highly Reduced Genomes of Protist Endosymbionts Show Evolutionary Convergence. *Current Biology* 2020; **30**: 925-933.e3.
15. Prokopchuk G, Korytář T, Juricová V, Majstorović J, Horák A, Šimek K, et al. Trophic flexibility of marine diplomonads - switching from osmotrophy to bacterivory. *ISME J* 2022; **16**: 1409–1419.
16. Seymour CO, Palmer M, Becraft ED, Stepanauskas R, Friel AD, Schulz F, et al. Hyperactive nanobacteria with host-dependent traits pervade Omnitrophota. *Nat Microbiol* 2023; **8**: 727–744.
17. Schön ME, Martijn J, Vosseberg J, Köstlbacher S, Ettema TJG. The evolutionary origin of host association in the Rickettsiales. *Nat Microbiol* 2022; **7**: 1189–1199.
18. Archibald JM. Endosymbiosis and eukaryotic cell evolution. *Current Biology* 2015; **25**: R911–R921.
19. Combes C. Parasitism: the ecology and evolution of intimate interactions. 2001. University of Chicago Press.
20. Brune A, Dietrich C. The gut microbiota of termites: digesting the diversity in the light of ecology and evolution. *Annu Rev Microbiol* 2015; **69**: 145–166.
21. Needham DM, Poirier C, Bachy C, George EE, Wilken S, Yung CCM, et al. The microbiome of a bacterivorous marine choanoflagellate contains a resource-demanding obligate bacterial associate. *Nat Microbiol* 2022; **7**: 1466–1479.
22. Castelli M, Nardi T, Gammuto L, Bellinzona G, Sabaneyeva E, Potekhin A, et al. Host association and intracellularity evolved multiple times independently in the Rickettsiales. *Nat Commun* 2024; **15**: 1093.
23. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 2013; **30**: 772–780.
24. Katoh K, Standley DM. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* 2016; **32**: 1933–1942.
25. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; **25**: 1972–1973.
26. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020; **37**: 1530–1534.
27. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; **14**: 587–589.

28. Wickham H. ggplot2: elegant graphics for data analysis. 2016. springer.
29. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017; **8**: 28–36.
30. Huang Y, Niu B, Gao Y, Fu L, Li W. CD-HIT Suite: a web server for clustering and comparing biological sequences. *Bioinformatics* 2010; **26**: 680–682.
31. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately Maximum-Likelihood Trees for Large Alignments. *PLoS One* 2010; **5**: e9490-.
32. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019; **47**: D309–D314.
33. Whelan S, Irisarri I, Burki F. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* 2018; **34**: 3929–3930.
34. Ali RH, Bogusz M, Whelan S. Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments. *Mol Biol Evol* 2019; **36**: 2340–2351.
35. Viklund J, Ettema TJG, Andersson SGE. Independent Genome Reduction and Phylogenetic Reclassification of the Oceanic SAR11 Clade. *Mol Biol Evol* 2012; **29**: 599–615.
36. Emms DM, Kelly S. OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 2019; **20**: 238.
37. Obiol A, Muhovic I, Massana R. Oceanic heterotrophic flagellates are dominated by a few widespread taxa. *Limnol Oceanogr* 2021; **66**: 4240–4253.
38. Wilken S, Yung CCM, Poirier C, Massana R, Jimenez V, Worden AZ. Choanoflagellates alongside diverse uncultured predatory protists consume the abundant open-ocean cyanobacterium Prochlorococcus. *Proceedings of the National Academy of Sciences* 2023; **120**: e2302388120.
39. Pachiadaki MG, Brown JM, Brown J, Bezuidt O, Berube PM, Biller SJ, et al. Charting the complexity of the marine microbiome through single-cell genomics. *Cell* 2019; **179**: 1623-1635.e11.
40. McCutcheon JP, Moran NA. Extreme genome reduction in symbiotic bacteria. *Nat Rev Microbiol* 2012; **10**: 13–26.
41. Lefoulon E, Clark T, Guerrero R, Cañizales I, Cardenas-Callirgos JM, Junker K, et al. Diminutive, degraded but dissimilar: Wolbachia genomes from filarial nematodes do not conform to a single paradigm. *Microb Genom* 2020; **6**.

42. Dunning Hotopp JC, Lin M, Madupu R, Crabtree J, Angiuoli S V, Eisen J, et al. Comparative Genomics of Emerging Human Ehrlichiosis Agents. *PLoS Genet* 2006; **2**: e21-.
43. Hemsley CM, O'Neill PA, Essex-Lopresti A, Norville IH, Atkins TP, Titball RW. Extensive genome analysis of *Coxiella burnetii* reveals limited evolution within genomic groups. *BMC Genomics* 2019; **20**: 441.
44. Mironov T, Sabaneyeva E. A robust symbiotic relationship between the ciliate *Paramecium multimicronucleatum* and the bacterium *Ca. Trichorickettsia mobilis*. *Front Microbiol* 2020; **11**: 603335.
45. Seshadri R, Samuel J. Genome Analysis of *Coxiella burnetii* Species: Insights into Pathogenesis and Evolution and Implications for Biodefense. *Ann N Y Acad Sci* 2005; **1063**: 442–450.
46. Brenner AE, Muñoz-Leal S, Sachan M, Labruna MB, Raghavan R. *Coxiella burnetii* and Related Tick Endosymbionts Evolved from Pathogenic Ancestors. *Genome Biol Evol* 2021; **13**: evab108.
47. Muñoz-Gómez SA, Hess S, Burger G, Lang BF, Susko E, Slamovits CH, et al. An updated phylogeny of the Alphaproteobacteria reveals that the parasitic Rickettsiales and Holosporales have independent origins. *Elife* 2019; **8**: e42535.
48. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* 2010; **59**: 307–321.
49. Parks DH, Chuvochina M, Rinke C, Mussig AJ, Chaumeil P-A, Hugenholtz P. GTDB: an ongoing census of bacterial and archaeal diversity through a phylogenetically consistent, rank normalized and complete genome-based taxonomy. *Nucleic Acids Res* 2022; **50**: D785–D794.
50. Yurchenko T, Ševčíková T, Příbyl P, El Karkouri K, Klimeš V, Amaral R, et al. A gene transfer event suggests a long-term partnership between eustigmatophyte algae and a novel lineage of endosymbiotic bacteria. *ISME J* 2018; **12**: 2163–2175.
51. Vannini C, Boscaro V, Ferrantini F, Benken KA, Mironov TI, Schweikert M, et al. Flagellar Movement in Two Bacteria of the Family Rickettsiaceae: A Re-Evaluation of Motility in an Evolutionary Perspective. *PLoS One* 2014; **9**: e87718-.
52. Martijn J, Schulz F, Zaremba-Niedzwiedzka K, Viklund J, Stepanauskas R, Andersson SGE, et al. Single-cell genomics of a rare environmental alphaproteobacterium provides unique insights into Rickettsiaceae evolution. *ISME J* 2015; **9**: 2373–2385.

53. Castelli M, Sabaneyeva E, Lanzoni O, Lebedeva N, Floriano AM, Gaiarsa S, et al. Deianiraea, an extracellular bacterium associated with the ciliate Paramecium, suggests an alternative scenario for the evolution of Rickettsiales. *ISME J* 2019; **13**: 2280–2294.
54. Su J, Jin L, Jiang Q, Sun W, Zhang F, Li Z. Phylogenetically Diverse ureC Genes and Their Expression Suggest the Urea Utilization by Bacterial Symbionts in Marine Sponge *Xestospongia testudinaria*. *PLoS One* 2013; **8**: e64848-.
55. Medina M, Collins AG, Taylor JW, Valentine JW, Lipps JH, Amaral-Zettler L, et al. Phylogeny of Opisthokonta and the evolution of multicellularity and complexity in Fungi and Metazoa. *Int J Astrobiol* 2003; **2**: 203–211.
56. Chen Y-H, Chiang P-W, Rogozin DY, Degermendezhy AG, Chiu H-H, Tang S-L. Salvaging high-quality genomes of microbial species from a meromictic lake using a hybrid sequencing approach. *Commun Biol* 2021; **4**: 996.
57. Woznica A. What choanoflagellates can teach us about symbiosis. *PLoS Biol* 2024; **22**: e3002561-.
58. Duron O, Doublet P, Vavre F, Bouchon D. The importance of revisiting Legionellales diversity. *Trends Parasitol* 2018; **34**: 1027–1037.
59. Plotnikov AO, Balkin AS, Gogoleva NE, Lanzoni O, Khlopko YA, Cherkasov S V, et al. High-Throughput Sequencing of the 16S rRNA Gene as a Survey to Analyze the Microbiomes of Free-Living Ciliates Paramecium. *Microb Ecol* 2019; **78**: 286–298.
60. Costa TRD, Patkowski JB, Macé K, Christie PJ, Waksman G. Structural and functional diversity of type IV secretion systems. *Nat Rev Microbiol* 2024; **22**: 170–185.
61. Haiwei L, Miklós C, L HA, Ann MM. Evolution of Divergent Life History Strategies in Marine Alphaproteobacteria. *mBio* 2013; **4**: 10.1128/mbio.00373-13.
62. Khan F, Tabassum N, Anand R, Kim Y-M. Motility of *Vibrio* spp.: regulation and controlling strategies. *Appl Microbiol Biotechnol* 2020; **104**: 8187–8208.
63. Dunbar SA, Winkler HH. Increased and controlled expression of the *Rickettsia prowazekii* ATP/ADP translocase and analysis of cysteine-less mutant translocase. *Microbiology (N Y)* 1997; **143**: 3661–3669.
64. Deeg CM, Zimmer MM, George EE, Husnik F, Keeling PJ, Suttle CA. Chromulinavorax destructans, a pathogen of microzooplankton that provides a window into the enigmatic candidate phylum Dependotia. *PLoS Pathog* 2019; **15**: e1007801-.
65. Kolo AO, Raghavan R. Impact of endosymbionts on tick physiology and fitness. *Parasitology* 2023; **150**: 859–865.

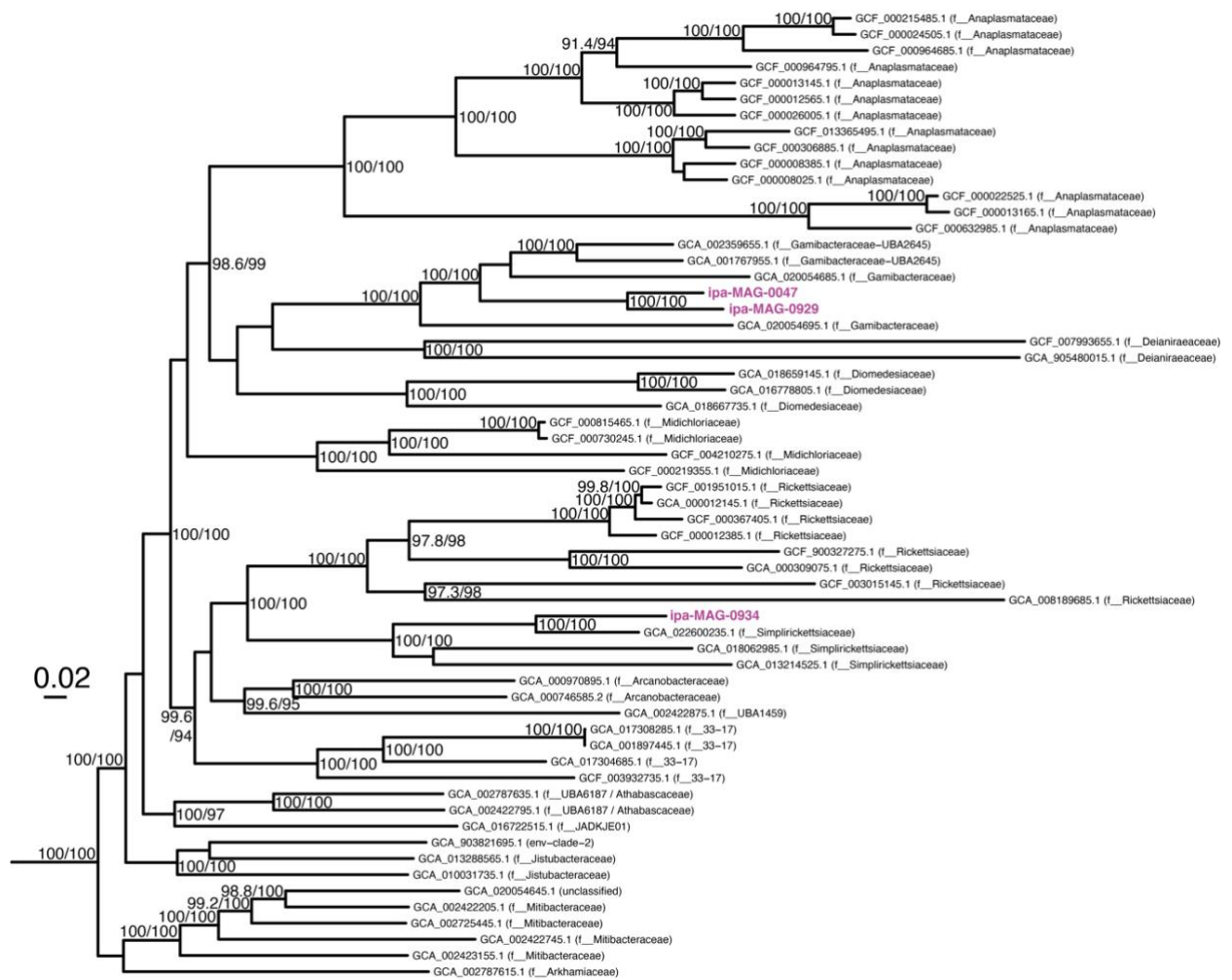
66. Mrowicka M, Mrowicki J, Dragan G, Majsterek I. The importance of thiamine (vitamin B1) in humans. *Biosci Rep* 2023; **43**: BSR20230374.
67. Paerl RW, Sundh J, Tan D, Svenningsen SL, Hylander S, Pinhassi J, et al. Prevalent reliance of bacterioplankton on exogenous vitamin B1 and precursor availability. *Proceedings of the National Academy of Sciences* 2018; **115**: E10447–E10456.
68. Paerl RW, Bertrand EM, Rowland E, Schatt P, Mehiri M, Niehaus TD, et al. Carboxythiazole is a key microbial nutrient currency and critical component of thiamin biosynthesis. *Sci Rep* 2018; **8**: 5940.
69. Hickin ML, Kakumanu ML, Schal C. Effects of Wolbachia elimination and B-vitamin supplementation on bed bug development and reproduction. *Sci Rep* 2022; **12**: 10270.
70. Schmitz-Esser S, Patrick T, Roland A, Jacqueline M, Michael W, Thomas R, et al. The Genome of the Amoeba Symbiont “Candidatus Amoebophilus asiaticus” Reveals Common Mechanisms for Host Cell Interaction among Amoeba-Associated Bacteria. *J Bacteriol* 2010; **192**: 1045–1057.
71. Jackson R, Monnin D, Patapiou PA, Golding G, Helanterä H, Oettler J, et al. Convergent evolution of a labile nutritional symbiosis in ants. *ISME J* 2022; **16**: 2114–2122.
72. Gerdol M, Sollitto M, Pallavicini A, Castellano I. The complex evolutionary history of sulfoxide synthase in ovolith biosynthesis. *Proceedings of the Royal Society B: Biological Sciences* 2019; **286**: 20191812.
73. Spies HSC, Steenkamp DJ. Thiols of Intracellular Pathogens. *Eur J Biochem* 1994; **224**: 203–213.
74. Brancaccio M, Tangherlini M, Danovaro R, Castellano I. Metabolic Adaptations to Marine Environments: Molecular Diversity and Evolution of Ovolith Biosynthesis in Bacteria. *Genome Biol Evol* 2021; **13**: evab169.
75. Wang J, Li Q, Huang L. Effect of the zinc transporter ZupT on the virulence mechanisms of mesophilic *Aeromonas salmonicida* SRW-OG1. *Animal Research and One Health* 2023; **1**: 30–42.
76. Cerasi M, Liu JZ, Ammendola S, Poe AJ, Petrarca P, Pesciaroli M, et al. The ZupT transporter plays an important role in zinc homeostasis and contributes to *Salmonella enterica* virulence†. *Metallomics* 2014; **6**: 845–853.
77. Moreau GB, Qin A, Mann BJ. Zinc Acquisition Mechanisms Differ between Environmental and Virulent *Francisella* Species. *J Bacteriol* 2018; **200**: 10.1128/jb.00587-17.

78. Zackular JP, J KR, A LC, N BW, Noel MC, J CW, et al. ZupT Facilitates *Clostridioides difficile* Resistance to Host-Mediated Nutritional Immunity. *mSphere* 2020; **5**: 10.1128/msphere.00061-20.
79. Martinez E, Cantet F, Fava L, Norville I, Bonazzi M. Identification of OmpA, a *Coxiella burnetii* protein involved in host cell invasion, by multi-phenotypic high-content screening. *PLoS Pathog* 2014; **10**: e1004013.
80. Park JS, Lee WC, Yeo KJ, Ryu K-S, Kumarasiri M, Heseck D, et al. Mechanism of anchoring of OmpA protein to the cell wall peptidoglycan of the gram-negative bacterial outer membrane. *The FASEB Journal* 2012; **26**: 219.
81. Leung KY, Wang Q, Yang Z, Siame BA. *Edwardsiella piscicida*: A versatile emerging pathogen of fish. *Virulence* 2019; **10**: 555–567.
82. Jin M, He J, Li J, Hu Y, Sun D, Gu H. *Edwardsiella piscicida* YccA: A novel virulence factor essential to membrane integrity, mobility, host infection, and host immune response. *Fish Shellfish Immunol* 2022; **126**: 318–326.
83. Chazan A, Das I, Fujiwara T, Murakoshi S, Rozenberg A, Molina-Márquez A, et al. Phototrophy by antenna-containing rhodopsin pumps in aquatic environments. *Nature* 2023; **615**: 535–540.
84. Balashov SP, Imasheva ES, Boichenko VA, Antón J, Wang JM, Lanyi JK. Xanthorhodopsin: a proton pump with a light-harvesting carotenoid antenna. *Science (1979)* 2005; **309**: 2061–2064.
85. Imasheva ES, Balashov SP, Choi AR, Jung K-H, Lanyi JK. Reconstitution of *Gloeobacter violaceus* rhodopsin with a light-harvesting carotenoid antenna. *Biochemistry* 2009; **48**: 10948–10955.
86. Qiu J, Luo Z-Q. *Legionella* and *Coxiella* effectors: strength in diversity and activity. *Nat Rev Microbiol* 2017; **15**: 591–605.
87. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences* 2019; **116**: 20574 LP – 20583.
88. Koonin E V, Krupovic M. New faces of prokaryotic mobile genetic elements: Guide RNAs link transposition with host defense mechanisms. *Curr Opin Syst Biol* 2023; **36**: 100473.
89. Zhong A, Jiang X, Hickman AB, Klier K, Teodoro GIC, Dyda F, et al. Toxic antiphage defense proteins inhibited by intragenic antitoxin proteins. *Proceedings of the National Academy of Sciences* 2023; **120**: e2307382120.

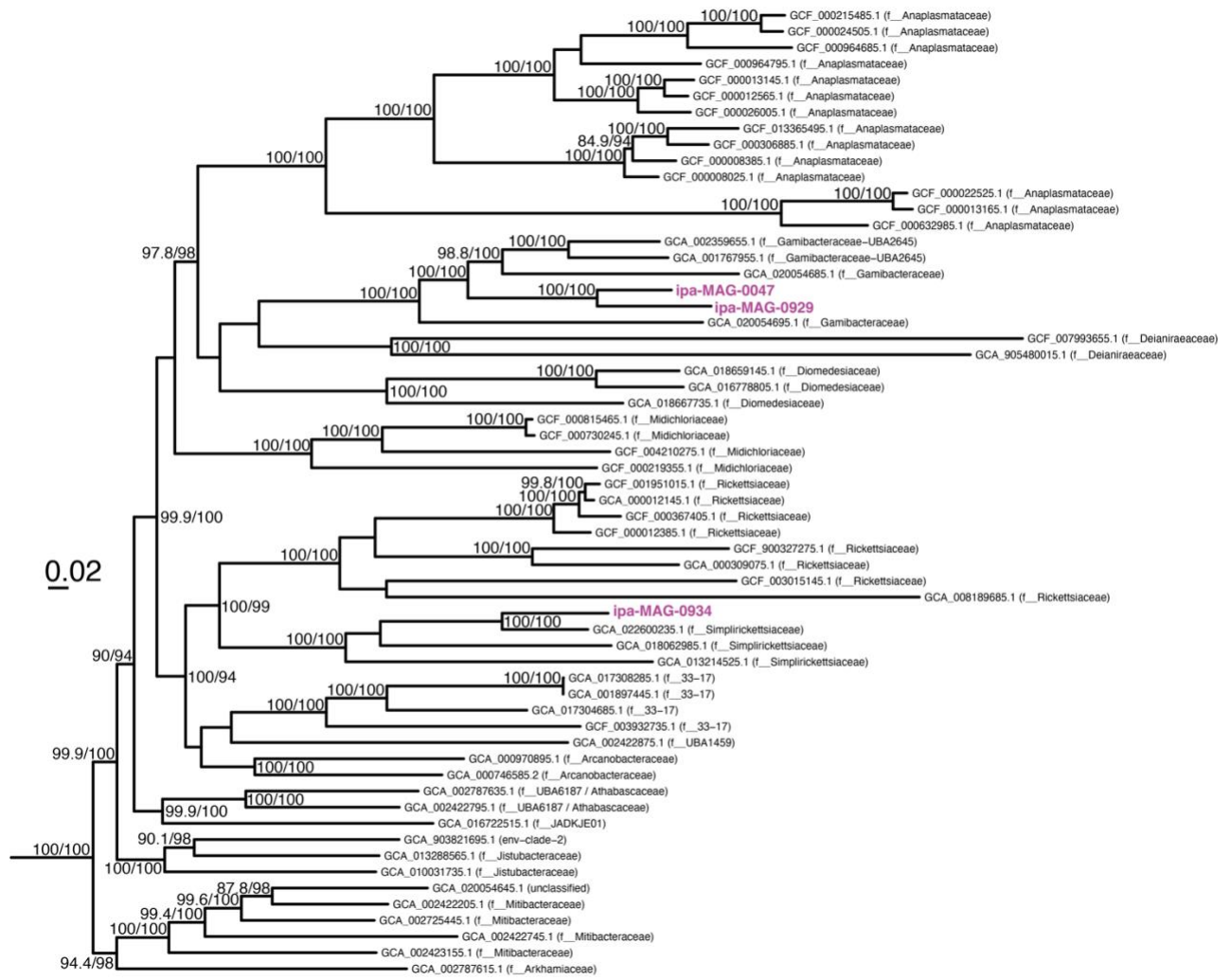
90. Wright GD, Thompson PR. Aminoglycoside phosphotransferases: proteins, structure, and mechanism. *Front Biosci* 1999; **4**: D9-21.
91. Foulquier E, Galinier A. YvcK, a protein required for cell wall integrity and optimal carbon source utilization, binds uridine diphosphate-sugars. *Sci Rep* 2017; **7**: 4139.
92. Nikoh N, Hosokawa T, Moriyama M, Oshima K, Hattori M, Fukatsu T. Evolutionary origin of insect–Wolbachia nutritional mutualism. *Proceedings of the National Academy of Sciences* 2014; **111**: 10257–10262.
93. Galindo LJ, Torruella G, Moreira D, Eglit Y, Simpson AGB, Völcker E, et al. Combined cultivation and single-cell approaches to the phylogenomics of nucleariid amoebae, close relatives of fungi. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2019; **374**: 20190094.
94. Ogata H, La Scola B, Audic S, Renesto P, Blanc G, Robert C, et al. Genome Sequence of *Rickettsia bellii* Illuminates the Role of Amoebae in Gene Exchanges between Intracellular Pathogens. *PLoS Genet* 2006; **2**: e76-.
95. Wernegreen JJ, Moran NA. Evidence for genetic drift in endosymbionts (Buchnera): analyses of protein-coding genes. *Mol Biol Evol* 1999; **16**: 83–97.
96. O SS, Yvonne B, S UM, Helge W. Characterization of the RND-Type Multidrug Efflux Pump MexAB-OprM of the Plant Pathogen *Pseudomonas syringae*. *Appl Environ Microbiol* 2008; **74**: 3387–3393.
97. P RS, L PJ, J MJ. The Rickettsial OmpB  $\beta$ -Peptide of *Rickettsia conorii* Is Sufficient To Facilitate Factor H-Mediated Serum Resistance. *Infect Immun* 2012; **80**: 2735–2743.
98. Sabrin A, Gioe BW, Gupta A, Grove A. An EmrB multidrug efflux pump in *Burkholderia thailandensis* with unexpected roles in antibiotic resistance. *Journal of Biological Chemistry* 2019; **294**: 1891–1903.
99. Huang Y, Jin F, Funato Y, Xu Z, Zhu W, Wang J, et al. Structural basis for the Mg<sup>2+</sup> recognition and regulation of the CorC Mg<sup>2+</sup> transporter. *Sci Adv* 2024; **7**: eabe6140.
100. Carroll RK, Rivera FE, Cavaco CK, Johnson GM, Martin D, Shaw LN. The lone S41 family C-terminal processing protease in *Staphylococcus aureus* is localized to the cell wall and contributes to virulence. *Microbiology (N Y)* 2014; **160**: 1737–1748.
101. Cirzi C, Dyckow J, Legrand C, Schott J, Guo W, Perez Hernandez D, et al. Queuosine-tRNA promotes sex-dependent learning and memory formation by maintaining codon-biased translation elongation speed. *EMBO J* 2023; **42**: e112507.

102. Díaz-Rullo J, González-Pastor JE. tRNA queuosine modification is involved in biofilm formation and virulence in bacteria. *Nucleic Acids Res* 2023; **51**: 9821–9837.
103. Liu B, Du H, Rutkowski R, Gartner A, Wang X. LAAT-1 Is the Lysosomal Lysine/Arginine Transporter That Maintains Amino Acid Homeostasis. *Science (1979)* 2012; **337**: 351–354.
104. Nikolaeva DD, Gelfand MS, Garushyants SK. Simplification of Ribosomes in Bacteria with Tiny Genomes. *Mol Biol Evol* 2021; **38**: 58–66.
105. Akanuma G, Ako K, Shota S, Fujio K, Yuh S, Satoru W, et al. Defect in the Formation of 70S Ribosomes Caused by Lack of Ribosomal Protein L34 Can Be Suppressed by Magnesium. *J Bacteriol* 2014; **196**: 3820–3830.
106. Sañudo-Wilhelmy SA, Cutter LS, Durazo R, Smail EA, Gómez-Consarnau L, Webb EA, et al. Multiple B-vitamin depletion in large areas of the coastal ocean. *Proceedings of the National Academy of Sciences* 2012; **109**: 14041–14045.
107. Fonseca C, Mendonça Filho JG, Reolid M, Duarte L V, de Oliveira AD, Souza JT, et al. First putative occurrence in the fossil record of choanoflagellates, the sister group of Metazoa. *Sci Rep* 2023; **13**: 1242.
108. V IE, Arielle W, Nicole K. Synergistic Cues from Diverse Bacteria Enhance Multicellular Development in a Choanoflagellate. *Appl Environ Microbiol* 2020; **86**: e02920-19.
109. Woznica A, Gerdt JP, Hulett RE, Clardy J, King N. Mating in the closest living relatives of animals is induced by a bacterial chondroitinase. *Cell* 2017; **170**: 1175–1183.
110. Blanco-Bercial L, Parsons R, Bolaños L, Johnson R, Giovannoni S, Curry R. The protist community mirrors seasonality and mesoscale hydrographic features in the oligotrophic Sargasso Sea. *Authorea* 2022.
111. Fraune S, Bosch TCG. Long-term maintenance of species-specific bacterial microbiota in the basal metazoan Hydra. *Proceedings of the National Academy of Sciences* 2007; **104**: 13146–13151.
112. Gong J, Qing Y, Zou S, Fu R, Su L, Zhang X, et al. Protist-Bacteria Associations: Gammaproteobacteria and Alphaproteobacteria Are Prevalent as Digestion-Resistant Bacteria in Ciliated Protozoa. *Front Microbiol* 2016; **7**.
113. Gillespie JJ, Nordberg EK, Azad AF, Sobral BWS. Phylogeny and Comparative Genomics: the Shifting Landscape in the Genomics Era. *Intracellular Pathogens II*. 2012. pp 84–141.

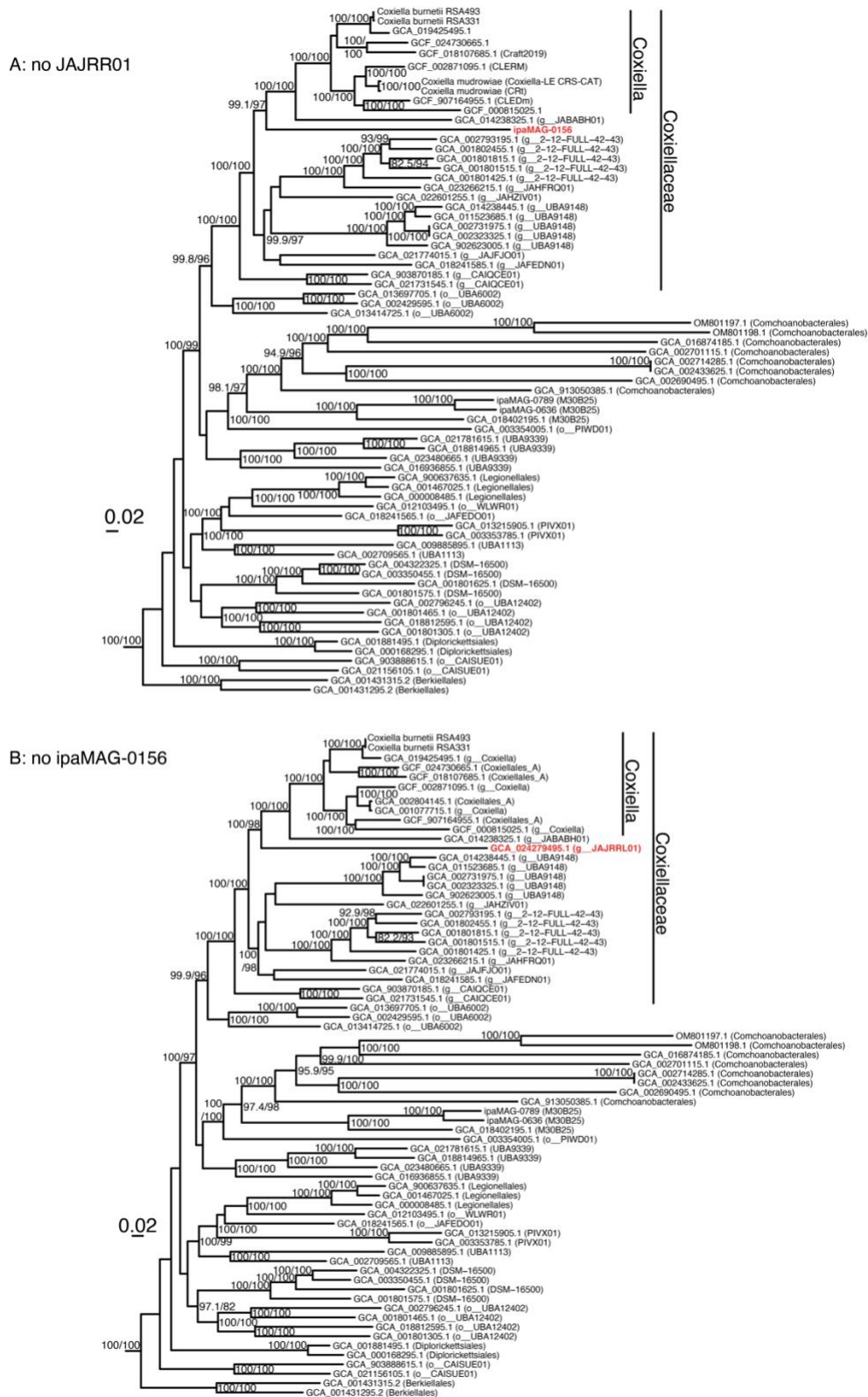
## Supplementary Figures



**Figure S1: Phylogenomic tree of Rickettsiales including *Ca. Arcanobacter lacustris*.** Maximum-likelihood reconstruction of Rickettsiales based on 79 concatenated single copy ortholog alignments, additionally including the MAG for *Ca. Arcanobacter lacustris*, which was not included in Figure 3 because it is low quality (45.29% estimated completeness, 1.9% contamination). Here, it is added (43 SCMG only, the rest demarked as missing positions) due to the critical phylogenomic position of the Arcanobacteraceae in this region of the tree and near to the newly identified Simplirickettsiaceae. Node support is based -on SH-aLRT [48] and ultra-fast bootstrap replicates and support is indicated for clades with SH-aLRT  $\geq 80\%$  and UFboot  $\geq 95\%$ .

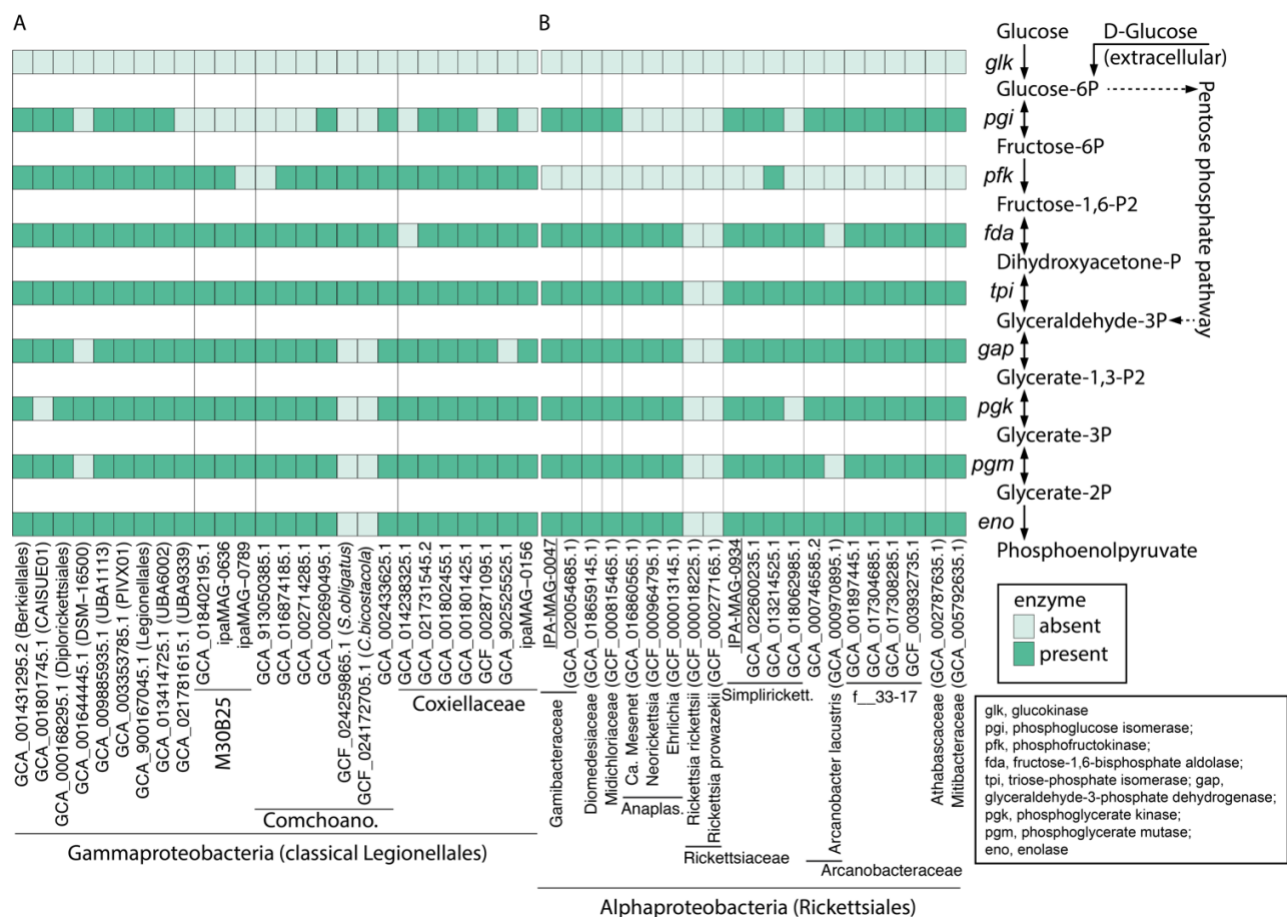


**Figure S2: Phylogenomic tree of Rickettsiales incorporating only SCMG present in the partial MAG for *Ca. Arcanobacter lacustris*.** Maximum-likelihood reconstruction of Rickettsiales (including *Ca. Arcanobacter lacustris*) based on 43 of 79 SCMG that were present in the incomplete *Ca. Arcanobacter lacustris* MAG. Node support is based on SH-aLRT [48] and ultra-fast bootstrap replicates and support is indicated for clades with SH-aLRT  $\geq 80\%$  and UFboot  $\geq 95\%$ .



**Figure S3: Phylogenomic trees of traditional Legionellales to investigate potential long-branch attraction in the Marinicoxiellaceae.** Phylogenomic trees as in Figure 4 were reconstructed independent after exclusion of (A) JAJRR01 and (B) ipaMAG-0156, respectively, to examine impact on topology of the Coccidiellales. The member of the Marinicoxiellaceae that was kept is highlighted in red. Node support is based on SH-aLRT [48] and ultra-fast bootstrap replicates and support is indicated for clades with SH-aLRT  $\geq 80\%$  and UFboot  $\geq 95\%$ .





**Figure S5: Reduction of Glycolysis pathway in classical Legionellales (A) and Rickettsiales (B).** (A)

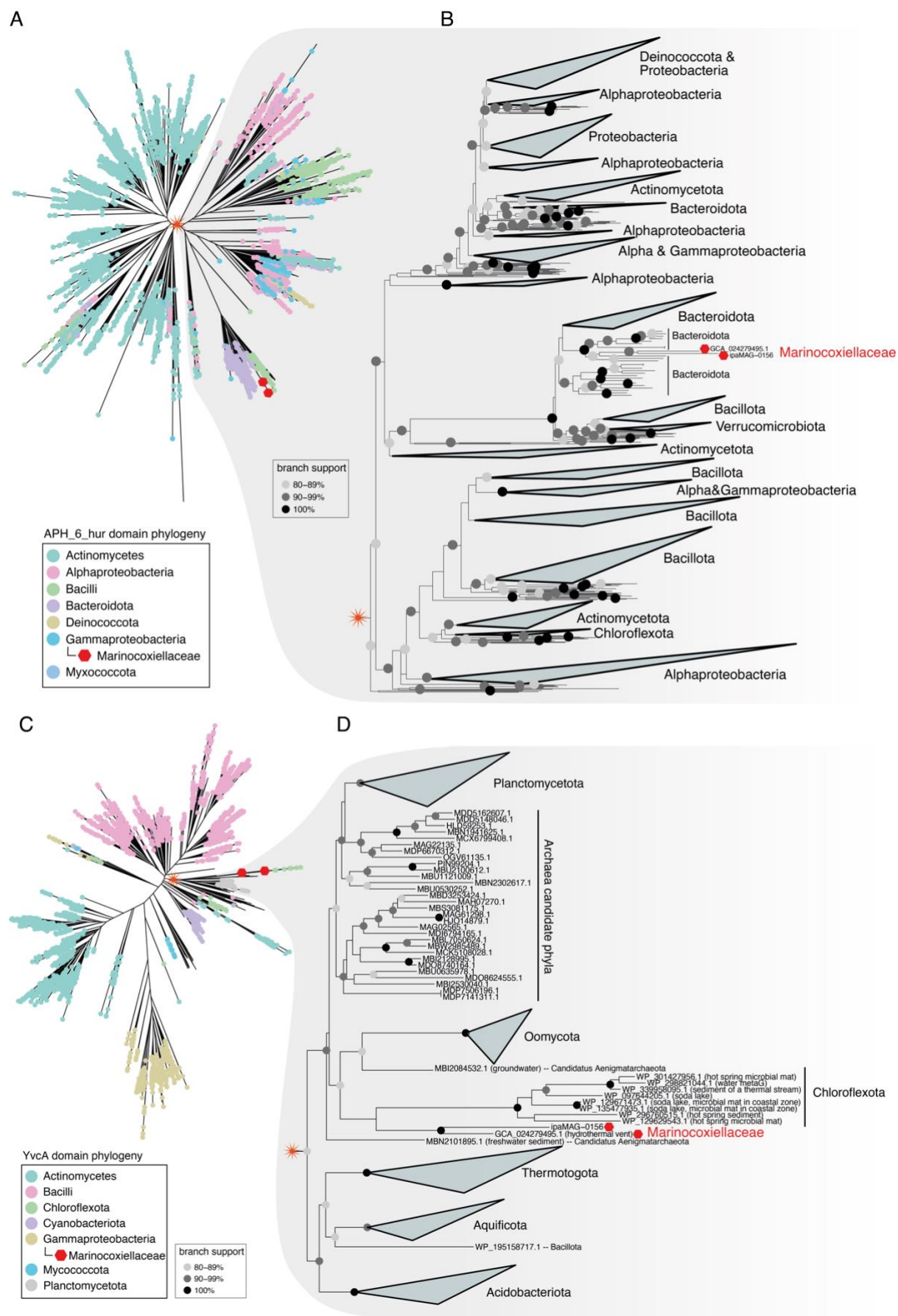
Glycolysis pathway completeness across traditional Legionellales, comparing representative members of families.

Annotations are based on KEGG orthology annotations associated with the Glycolysis (Embden-Meyerhof

pathway) module (M00001). (B) Glycolysis pathway completeness across Rickettsiales, comparing

representative members of families. Note that Rickettsiales are still able to conduct glycolysis through a

connection to the Entner-Doudoroff glycolysis pathway when lacking phosphofructokinase (pgi).



**Figure S6:** (see figure legend on next page)

**Figure S6: Horizontal acquisition APH\_6\_hur domain and YvcA domain in Marinocoxiellaceae.** Maximum-likelihood-based phylogenies of APH\_6\_hur (A) and YvcA (C) based on representative sequences from NCBI protein models NF016533.0 and TIGR01826.1, respectively, as well as blastp homologs of the Marinocoxiellaceae query sequences. Taxonomy information is based on NCBI taxon IDs and UFboot support values are only shown for the zoomed sections of the phylogenies containing Marinocoxiellaceae family sequences for visualization purposes. Phylogenies were reconstructed using 5184 sequences with 255 positions (APH\_6\_hur) and 3383 sequences and 307 positions (YvcA) respectively, using FastTree with model settings ‘-lg -cat 3 - gamma’. Node support is based on ultra-fast bootstrap replicates and support is indicated for clades with UFboot  $\geq 80\%$ .

## Chapter 4

### Non-photosynthetic siblings of cyanobacteria associate with eukaryotes in the open ocean

---

Fabian Wittmers<sup>1,2</sup>, Jacqui Comstock<sup>3</sup>, Camille Poirier<sup>2</sup>, David M. Needham<sup>2</sup>, Rex Malstrom<sup>4</sup>, Craig A. Carlson<sup>5</sup>, Alexandra Z. Worden<sup>1,2,\*</sup>

<sup>1</sup>Marine Biological Laboratory, Woods Hole, MA 02543, USA

<sup>2</sup>Ocean EcoSystems Biology Unit, RD3, GEOMAR Helmholtz Centre for Ocean Research, 24148 Kiel, DE

<sup>3</sup>Marine Science Institute, Department of Ecology, Evolution and Marine Biology, University of California Santa Barbara, 93117 Santa Barbara, CA, US

<sup>4</sup>DOE Joint Genome Institute, 94720 Berkeley, CA, US

\* For correspondence: A.Z. Worden, Marine Biological Laboratory, Woods Hole, MA, USA.  
Email: [azworden@mbl.edu](mailto:azworden@mbl.edu)

This chapter has been submitted as an article manuscript to the journal *Current Biology* in July 2024 and is currently (August 15<sup>th</sup>, 2024) in revision.

## **Abstract**

Margulisbacteria are elusive uncultivated bacteria that have illuminated potential evolutionary transitions in the progenitor of Cyanobacteria, the latter being a critically important phylum that underpins all known oxygenic photosynthesis. Members of the non-photosynthetic candidate phylum Ca. Margulisbacteria were first discovered in a sulfidic spring and since in other habitats, including freshwater environments and the deep sea. Currently, the phylum partitions into three groups, the Riflemargulisbacteria, largely from sediments and groundwater, the Termititenax from insect gut microbiomes, and the Marinamargulisbacteria, from marine samples. We noted an unusual distribution of Marinamargulisbacteria amplicons in size-fractionated seawater samples from the sunlit euphotic zone and the deep twilight zone below. Further, our sequencing of wild marine protist single cells rendered genomic information for three distinct marinamargulisbacterial lineages co-associated with uncultivated, non-photosynthetic Stramenopila and Opisthokonta protists. Phylogenomic analyses combining these data and available Metagenome-assembled genomes (MAGs) and single amplified genomes (SAGs) from sorted bacteria revealed new lineages within the Marinamargulisbacteria. These lineages appear to delineate by their environment, forming clades comprising freshwater, marine pelagic, or sediments / hypoxic taxa. The remarkable diversity of Margulisbacteria indicates a high rate of success in colonizing various habitats, potentially in a conserved strategy involving eukaryotic hosts.

## Introduction

The Margulisbacteria were initially identified as a unique group (ZB3) through the analysis of environmental 16S rRNA gene sequences [1]. However, it was only through phylogenomic analysis of MAGs and SAGs that their close relationship with Cyanobacteria was realized [2, 3], alongside several other non-photosynthetic groups that are related to Cyanobacteria. The Melainabacteria and Sericytochromatia form sister groups of the Cyanobacteria, while the Saganbacteria (formerly WOR-1) and Margulisbacteria (formerly ZB3 or RBX1, depending on the study) comprise two adjacent candidate phyla related to the latter trio [2–4].

Collectively, these lineages possess traits of the ancestral lineage that gave rise to Cyanobacteria and help trace the evolution of aerobic respiration, an important step connected to the development of oxygenic photosynthesis [3, 5]. Marine Margulisbacteria, specifically, encode an electron transport chain that could support aerobic growth of the lineage [3].

The Margulisbacteria remain uncultivated; thus, relatively little is known about their niches. This is especially true where the only genomic information available comes from bulk collection of water. In such cases many litres are filtered, and the captured biomass is sequenced together, prohibiting assembly of genomes from micro-diverse or rare microbes and obfuscating the detection of cell-to-cell interactions.

## Material & Methods

### Single cell sorting, amplification, and sequencing

Samples were collected from the upper photic zone during multiple sampling campaigns in Monterey Bay and the North Pacific Ocean between 2014 and 2019. Sorting was performed using freshly collected samples that had been passed through a 30  $\mu\text{m}$  nylon mesh upon collection and subsequently concentrated by gravity over a 0.8  $\mu\text{m}$  Supor filter (Pall Corporation). Single cells were sorted from environmental samples collected from the North Pacific (Monterey Bay). An InFlux high speed sorter (BD) equipped with 488 nm and 355 nm lasers and a 70  $\mu\text{m}$  nozzle, operating on the high purity Single-Cell sorting mode, with sterile nuclease-free 1x PBS as sheath fluid (ThermoFisher cat# AM9625 and AM9932), was used in all cases. First, samples were stained for 10-20 min in the dark with LysoTracker Green DND-26 (25 nM, final concentration). Unstained controls were run for all samples such that positive signal from the stain was always discerned as compared to the control. Forward Angle Light Scatter (FALS), and green fluorescence from LysoTracker (520/35 nm bandpass) were used as selection criteria, while all other particles were selected for exclusion. The drop delay and sort quality were regularly controlled by sorting a known number of fluorescent bead standards and counting them by epifluorescence microscopy. Immediately prior to sorting plates were illuminated with UV irradiation for 2 min. Negative and positive controls (for subsequent genome amplification reactions) were performed on all plates by leaving a subset of wells empty, and by having a column of wells receiving multiple cells within each 384 well plate, respectively. After sorting, plates were covered with foil seals and frozen at  $-80^{\circ}\text{C}$ . Single-protist cells were subjected to Multiple Displacement Amplification using the repliG or WGA-X enzyme, as previously described [6]. Sequencing was performed on the Illumina platform.

### Recovery of Margulisbacteria assemblies

Demultiplexed paired-end reads were quality controlled and trimmed using trimmomatics (v0.39; [7]), with a trimming cutoff quality score  $<3$  or when a 25 bp sliding window average was  $<30$ . Assembly was performed using SPAdes (v3.15.5; [8]) with the single cell option ``-sc`` and kmers 21, 33, 55, 77, 99 and 127, considering both paired and additional unpaired reads that passed quality-control. All assembled contigs were filtered and only those  $>1\text{kb}$  were retained. Sequence alignment maps for filtered contigs were generated using bowtie2

(v2.4.1; [9]), with default setting and unaligned reads removed from the alignment map. Sequence alignment files were sorted and compressed using samtools (v1.6; [10]) ‘view -bS’, ‘sort’ and ‘index’. Complete or fragments of 16S or 18S rRNA genes were identified using checkM’s (v1.2.2; [11]) subcommand ‘ssu\_finder’ with default settings. Sorted eukaryotes were classified based on recovered 18S rRNA gene sequences and their top blastn (v2.14; [12]) homology against SILVA (v138.1; [13]).

Prokaryotic bins were recovered using Anvi’o (v8; [14]), using ‘anvi-profile’ and >95% nucleotide identity. Cellular single copy marker genes for interactive genome completeness estimate were identified using ‘anvi-run-hmms’ and the default database. Preliminary taxonomy of the recovered genes was predicted at the protein level using ‘anvi-run-scg-taxonomy’ with default settings. Additionally, ‘anvi-run-hmms’ was run using a set of 47 HMMs for marker genes of eukaryotic giant viruses (NCVOGs; [6]). Profiles were manually inspected for anomalies in GC% content and tetranucleotide frequency and inspected interactively, binning potential prokaryotic genomes. Eukaryotic single cell sort assemblies comprised multiple clusters of contigs with distinct GC-content far exceeding typical variation observed within a single genome. Although especially genomic and organelle-derived DNA differs in GC-content, some of these distinct GC-regions far exceeded the genome size of organelles and were instead comprised of clusters of distinct prokaryotic DNA. All bins of GC anomalies were subsequently screened for contaminating viral sequences (both eukaryotic viruses from the NCLDV family and phages) using the following approach: Proteins were queried against NCBI nr using diamond blastp (v2.14; [12]) with E-value <  $1 \times 10^{-04}$  and output format 6, including an additional column for subject NCBI taxonomy ID. IDs were converted into taxonomy strings using NCBI taxid2lineage information (obtained through: [github.com/zyxue/ncbitax2lin](https://github.com/zyxue/ncbitax2lin)). Up to 50 query hits per protein were considered and hits were summed up for each contig individually when calculating the percent of viral hits. Contigs were classified as viral and removed from the raw bin when at least 2% of hits were viral taxa.

Taxonomy of putatively prokaryotic bins was predicted using GTDB-Tk’s (v1.0.2; [15]) ‘identify’, ‘align’ and ‘classify’ consecutively, with GTDB version 214.1 as the reference database and default settings when meeting GTDB-Tk minimum classification cutoffs. Viral clean bins were quality-assessed using checkM’s (v1.2.2; [6]) ‘lineage\_wf’ subcommand with default settings. The lack of high-quality or near-complete *Marinamargulisbacteria* genomes from the eukaryotic SAGs indicates that there were likely not multiple clonal *Marinamargulisbacteria* cells associated with one protistan cell, something observed

previously for a putatively intracellular symbiont of *B. minor* [16]; see also chapter 3. We were also unable to detect any prey organisms, such as *Prochlorococcus*, which the three heterotrophic protists herein are known predators of [17], suggesting that predation on Margulisbacteria is unlikely to cause the physical co-association sorted herein.

### Phylogenomics

Phylogenomic reconstruction of Margulisbacteria is based on an ortholog dataset build for Cyanobacteria (taxID: 1117) by eggNOG (v5.0; [18]), because of its evolutionary proximity to Margulisbacteria, for which themselves no designated ortholog-dataset is available. Margulisbacteria taxon sampling was based on all representative assemblies classified as members of the Margulisbacteria phylum by GTDB. Assemblies GCA\_003265865.1 (AG-333-B06), GCA\_003265905.1 (AG-439-L15), GCA\_003265925.1 (AG-414-C22), GCA\_003265815.1 (AG-343-K17), and GCA\_003265905.1 (AG-439-L15) used in previous reconstructions of the Margulisbacteria [3, 19] were not considered herein as they failed GTDB quality checks. Thus, several MAGs/SAGS analyzed in prior studies were excluded and we were able to effectively replace them with related higher quality MAGs/SAGs that are now available. Orthologs were identified using hmmscan (E-value  $<1 \times 10^{-30}$ ) against the prebuilt eggNOG hmm database for the respective lineage to identify proteins with single-copy orthologs (SCOs) in  $\geq 86\%$  of the protein catalog of the genome collection. SCOs were then dereplicated to ensure that a protein of a respective assembly was only part of one SCOs, to avoid the incorporation of the same protein into multiple later concatenated alignments. Highly discordant SCOs were removed by performing single protein reconstructions using IQ-TREE (v2.2.6; options: -B 1000 -bnni -m MFP) and alignments generated in MAFFT E-INS-i (v7.520; --auto; [20]) that had been trimmed via trimAl with the --gappyout option (v1.4.rev15; [21]). For each resulting tree bipartition counts were generated using 'tre\_make\_splits.pl' using IQ-TREE bootstrapping output in scripts obtained from 'github.com/novigit/broCode', followed by an all-vs-all pairwise computation of discordance using 'tre\_discordance\_two.pl'. The 10% most discordant SCOs based on average discordance against all other SCOs were removed resulting in a final set of 72 SCOs used for subsequent phylogenomic reconstruction.

For the lineage-specific phylogenomic reconstructions based on the 72 SCOs identified herein, non-homologous sites were masked for each SCO by pre-alignment of sequences using prequal (v1.02; [22, 23]), aligned using MAFFT E-INS-i as described above. SCO

alignments were then trimmed and ‘divvied’ using Divvier (v1.01; [23]) removing columns with more than 60% gaps and additionally filtering with ‘-partial’. Separate SCO alignments were concatenated using ‘catfasta2phym.pl –concatenate’ (github.com/113ayesian/catfasta2phym). The untreated concatenated alignment was further trimmed to remove compositionally heterogeneous sites using ‘alignment\_pruner.pl –chi2\_prune f0.01’, to filter out the top 1% most heterogeneous sites iteratively as established in [24] with 100 repetitions respectively. After iterative site removal, no taxa deviating significantly in a  $\chi^2$ -test. Maximum likelihood phylogenies were reconstructed with 9872 positions using IQ-TREE with 1000 ultra-fast bootstrap iterations and SH-aLRT (-alrt 1000) [25] under the Q.yeast+I+R7 model as selected by ModelFinder [26] within IQ-TREE. Accessions and metadata on assemblies included in the phylogenomic tree are available on figshare (see data availability statement).

### 16S rRNA gene phylogeny

Taxon sampling of Margulisbacteria was based on blastn searches of 16S rRNA gene sequences recovered herein against NCBI nr, incorporating all (near) full-length 16S rRNA gene hits up until we identified Riflemargulisbacteria sequences. The 16S rRNA gene sequence collection was updated by recovering 16S rRNA gene sequences from Margulisbacteria assemblies in GTDB using ‘checkM ssu\_finder’ as described above. When encoding multiple 16S rRNA gene sequences, the longest sequence was considered for the alignment. Details on sequences included in the phylogeny as well as metadata and corresponding references are available on figshare (see data availability statement). Selected 16S rRNA gene sequences were aligned using the MAFFT (v7.310; [20]) L-INS-i algorithm to avoid over-alignment [27] and subsequently trimmed using trimAl (v1.4.1; [21]). The trimmed alignment was further manually inspected for anomalies and curated, removing chimeric and incomplete sequences. Phylogenetic trees for both lineages were reconstructed using IQ-TREE (v.1.5.4; [28]) with ultra-fast bootstrap (-bb 1000) and SH-aLRT (-alrt 1000) with an optimized substitution model (TVMe+I+R4) based on ModelFinder’s [26] Bayesian information criterion. Phylogenetic trees and accessory data were visualized using ggplot2 (v3.4.2; [29]) with the ggtree extension (v3.6.1; [30]).

#### V4 16S rRNA gene amplicon analysis

We analyzed a total of 195 size-fractionated filters generated *in-situ* from two cruises carried out in the summer and autumn of 2021 at the Bermuda Atlantic Timeseries Study (BATS) site. Amplicon sequence variants (ASVs) were generated using DADA2 [31] with default settings. Taxonomy was predicted using the ‘qiime feature-classifier classify-sklearn’ function in qiime2 (v2020.11) and SILVA (v138) with default settings. Size-fractionated samples were classified into upper euphotic (samples above the DCM layer), the DCM layer, and the twilight zone (<0.1% incident light level to 1000 m; herein the deepest sample was 500 m).

#### Data Availability

Raw reads for single-cell sorts are available via NCBI BioProject PRJNA1061664, SRA accessions SRR27444404, SRR27444869, SRR27444848, SRR27444839, SRR27444953, and SRR27445056. Amplicon data is deposited in NCBI under BioProject PRJNA1066814. Full and trimmed alignments for phylogenetic and phylogenomic trees of Margulisbacteria, as well as all recovered bins and their predicted proteins are available through FigShare ([doi.org/10.6084/m9.figshare.c.7324622](https://doi.org/10.6084/m9.figshare.c.7324622)). Accessions of assemblies and nucleotide sequences used for phylogenies are available through supplementary tables S1 and S2 in FigShare.

## Results & Discussion

We examined the distribution of *Marinamargulisbacteria* in V4 16S rRNA gene amplicons from the Sargasso Sea collected in the summer (July) and autumn (November) of 2021. Specifically, we searched sequencing data from size-fractionated samples from the Bermuda Atlantic Time-series Study (BATS) site to gain insights into cell size and potential co-association of bacteria with larger particles. Our analyses showed that the *Marinamargulisbacteria* were present from the photic zone through 500 m at low relative abundances, but similar frequencies, in size fractions ranging from 0.2 – 1.2  $\mu\text{m}$ , where free-living bacteria like SAR11 and *Prochlorococcus* dominate, to 1.2 – 5  $\mu\text{m}$ , 5 – 20  $\mu\text{m}$  and >20  $\mu\text{m}$  where particles and living protistan cells are also captured (Figure 1A). The discovery of *Marinamargulisbacteria* throughout the four size fractions sampled suggests they have a free-living life stage, consistent with recovery of *Marinamargulisbacteria* SAGs from sorted bacterial cells in Atlantic waters [3], and a particle/cell associated stage, or facultatively occupy both niches.

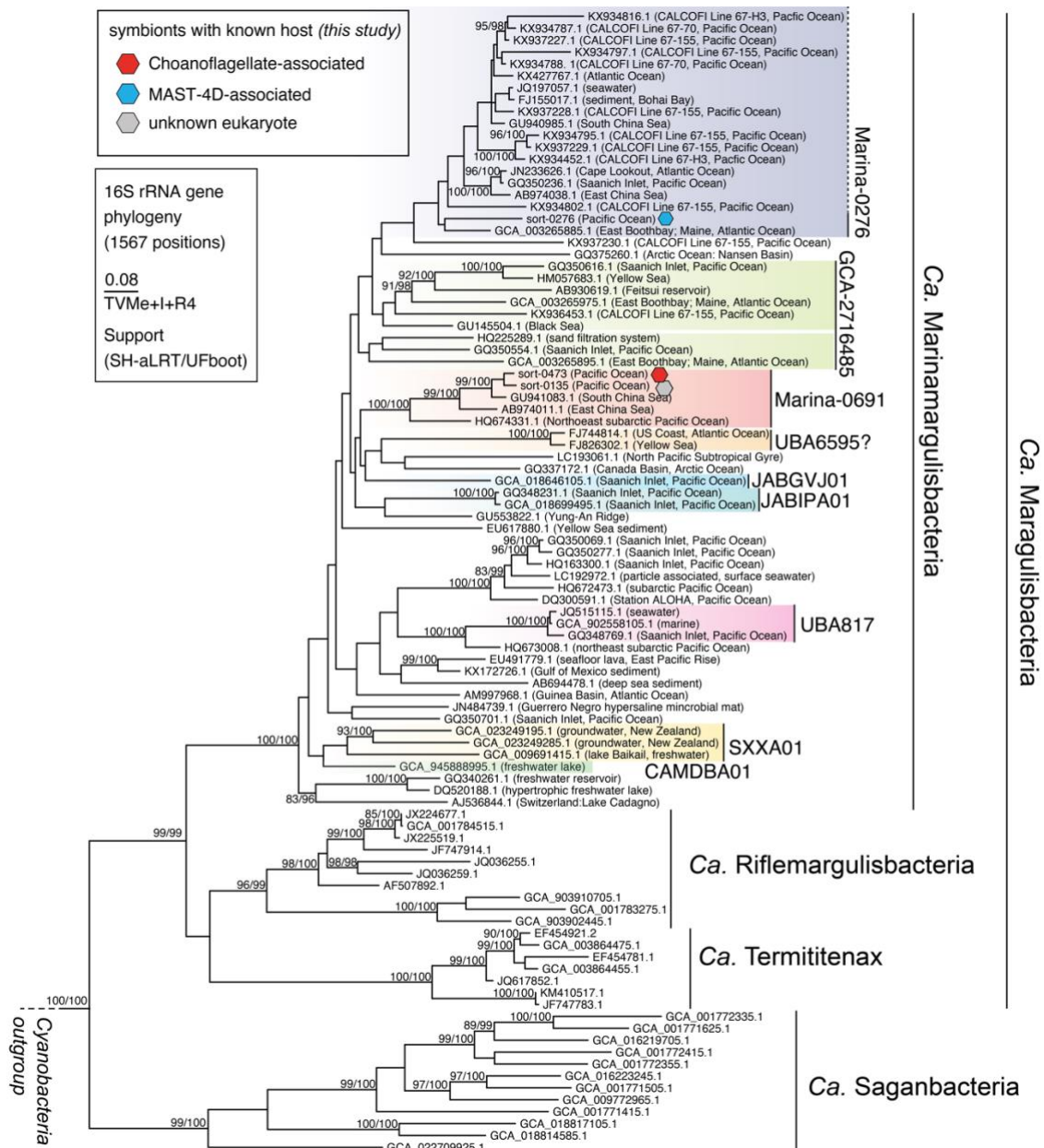
Two studies have pointed to a host-associated lifestyle for some *Margulisbacteria*. The first discovered the *Termititenax* class (Figure 1B), which exhibit complex microbiome interactions in termite guts [4]. These tiny *Margulisbacteria* are physically attached to ectosymbiotic (spirochete) bacteria that resided on heterotrophic Excavata protists that live within the guts of these insects [4]. The second study examined intracellular symbionts living in the placozoan *Trichoplax*, a simple marine animal, uncovering an uncultivated *Marinamargulisbacterium* (*Ruthmannia eludens*) that localizes to *Trichoplax* epithelial cells [19]. More recently, *Margulisbacteria* have been reported to reside in gill tissues of oysters [32]. Our results indicating a possible co-association of *Marinamargulisbacteria* not only with multicellular animals, but with protists, consistent with these earlier studies that point to co-associations with eukaryotes, even if some are from outside the *marinamargulisbacterial* class. These protists would be captured in the larger size-fractions sampled at BATS. Thus, we hypothesized that relationships based on physical co-association occur between members of *Marinamargulisbacteria* and unicellular eukaryotes in the pelagic ocean.



To examine our hypothesis, we employed fluorescence activated cell sorting (FACS) that targeted individual cells of non-photosynthetic protists in the ocean photic zone. Distinct margulisbacterial assemblies were recovered from multiple single cell sorts of heterotrophic protists and used to examine evolutionary relationships within the phylum, alongside MAGs and SAGs recruited from publicly available data (Figure 1B). Overall, the topology of our phylogenomic reconstruction supports prior work that established Saganbacteria and Margulisbacteria as separate phyla [2, 3], in contrast to some other studies. Further, our analyses placed the *Termititenax* from insect gut microbiomes in a sister relationship with the sediment associated fermentative *Riflemargulisbacteria*, adjacent to the putatively aerobic *Marinamargulisbacteria*.

Our phylogenomic analyses resolved distinct *Marinamargulisbacteria* lineages affiliated with the FACS-sorted protists (Figure 1B). Sort-0276 was a stramenopile belonging to the widespread marine MAST-4D protists [17, 33]. The co-associated Margulisbacterium was placed within a statistically supported clade that included pelagic MAGs or SAGs from several oceans, with a coastal Atlantic SAG (AG-343-D04) branching as its sister (Figure 1B). Partitioned in an adjacent position was an unsupported clade that included AAA071-K20 and AG-410-N11. The latter two MAGs formed one of the marine groups resolved by Matheus-Carnevali and colleagues [3], and purportedly branch within and basal to *R. eludans* [19] (the relationship with *R. eludans* was not examined because the genomic sequence is not available in public databases). Prior work also suggested that UBA6595 and UBA817 form individual clades [3], which our reconstruction agrees with, although we didn't observe statistical support for this node.

The other *Marinamargulisbacteria* captured in the protist sorts belonged to distinct groups. One branched deep within UBA817 (sort-0173, Figure 1B), a group that contains MAGs and SAGs from marine and estuarine water samples. It was co-associated with MAST-1C, another widespread lineage of heterotrophic stramenopile protists [17, 33]. Both MAST-1 and MAST-4 are single-cell heterotrophic **MA**rine **ST**ramenopiles, but evolutionarily positioned in two different stramenopiles subgroups, the Sagenista (MAST-4) and Pseudofungi (MAST-1) [34]. Size-fractionated data suggests that MAST-4 are picoplanktonic protists (0.8-3µm), while MAST-1 cells are larger components, found within the nanoplankton (3-20µm) [35]. Both MAST-1 and MAST-4 are among the most abundant marine stramenopile lineages known, with widespread global distributions in pelagic ecosystems, with the exception that MAST-4 is absent in polar systems, likely due to a limit in temperature tolerance [34, 36].



**Figure 2: Phylogenetic reconstruction of Margulisbacteria 16S rRNA gene sequences.** Maximum-likelihood reconstruction based on near full-length (including the sequences from single-cell sorts) and near full-length 16S rRNA gene sequences. The trimmed alignment with 1567 positions was evaluated based on the optimal selected model TVMe+I+R4. Node support is based on Shimodaira-Hasegawa approximate likelihood-ratio test (SH-aLRT) and 1000 ultra-fast bootstrap (UFboot) replicates and is indicated for clades with SH-aLRT  $\geq 80\%$  and UFboot  $\geq 95\%$ . The tree was rooted using diverse Cyanobacteria as an outgroup, with the outgroup being omitted in visualization. Select Marinamargulisbacteria clades are colored based on order-level taxonomy using GTDB naming when 16S rRNA gene sequences stemmed from genome assemblies. Hexagon color indicates taxonomy of the heterotrophic protist SAG the 16S rRNA gene sequence was recovered from. The connection of UBA6595 identified in the phylogenomic analysis with 16S sequences was determined using a 737 nt fragment of the 16S rRNA gene recovered from Caspian Sea MAG JAQCBC01.1 (GCA\_027621615.1) from 15 m depth [37]. V4 18S sampling at the same site confirmed the presence of both MAST clades and choanoflagellates at BATS [38].

The sort-0173 taxon was closest to a MAG from the BATS photic zone (GCA\_902558105.1, Figure 1B), where MAST-1C as well as MAST-4D are ubiquitous [38]. The third group was co-associated with choanoflagellates, Opisthokonts that are the closest living unicellular relatives of animals [17]. This group formed a deep-branching lineage distinct from previously reported marinimargulisbacterial clades. In addition to resolving these protist-associated marine groups, our analyses revealed several as-of-yet unrecognized non-marine clades within the ‘Marinamargulisbacteria’, including freshwater lineages (CAMDBA01, SXXA01, SXXA01\_A), and two from hypoxic seawater samples (JABIPA01, JABGVJ01) (Figure 1B).

To date, only four high-quality MAGs (average 91.5% complete) have been recovered from the Marinamargulisbacteria, as determined through counts of universally conserved proteins. Two of these are from freshwater clade SXXA01 and two deep branching UBA817 members (Table S2). However, some so-called universal proteins are known to be systematically absent from characterized symbionts [16]. The lack of completion for most marinamargulisbacterial MAGs and SAGs could therefore reflect some level of gene loss that should be characterized in future studies. Our assemblies did not reach completion levels that support reliable assessment of gene loss. Finally, using the 16S rRNA gene sequences recovered in our MAGs and near full-length cloned environmental 16S rRNA gene sequences, we could further establish that the three protist-associated lineages identified herein are present in the Pacific and Atlantic Oceans, as well as the China Sea (Figure 2). Further, we show that UBA6595, previously represented by just a single MAG from near South Africa in prior phylogenomic work [3], is present in the Caspian Sea, the Atlantic and Pacific (Figure 1B, Figure 2).

## Conclusions

Our results indicate that members of three (Marina-0691, Marina-0276, UBA817) of five marine Margulisbacteria clades are likely to have a lifestyle associated with protists. This suggests they are facultative ecto- or endosymbionts of eukaryotes. While little is known about the lifestyle of any of the non-photosynthetic sibling lineages of the Cyanobacteria, our findings combined with those from Margulisbacteria characterized in the multicellular Opisthokonts Placozoa, Oyster, and Insecta, underscore a predilection for association with eukaryotes. We postulate that the Margulisbacteria expose traits of the ancestral lineage that ultimately gave rise to Cyanobacteria as symbionts, and ultimately organelles, of eukaryotes.

Finally, co-association of three distinct lineages of marine Marinamargulisbacteria with pelagic heterotrophic protists, and continuous presence over sunlit and dark regions of the water column, provide first insights into their interconnected roles in ocean ecosystems.

### **Acknowledgements**

This work was supported by Simons Foundation International's BIOS-SCOPE Program as well as GBMF3877. We also thank the Monterey Bay Aquarium Research Institute for support, including the Captain and Crew of the R/V *Western Flyer*. We thank the officers and crew of the R/V *Atlantic Explorer*, R. Parsons for cruise and logistical support and the rest of the BIOS-SCOPE team for assisting in sampling of size-fractioned amplicon data. We are grateful to Worden Lab members who facilitated cruise work and especially S. Sudek, S. Wilken, and M. Hamilton.

### **Author Contributions**

F.W. and A.Z.W. conceptualized the analysis. L.S., C.P. and D.M.N. prepared sequencing plates. R.M. amplified DNA and facilitated sequencing. C.A.C., F.W. and J.C. collected size-fractioned DNA samples. J.C. extracted size-fractioned DNA, prepared amplicon sequencing plates, and created the initial feature table. F.W. performed taxonomic annotations of size-fractioned amplicon data and data analysis of single cell genomes.

## References

1. Elshahed MS, Senko JM, Najar FZ, Kenton SM, Roe BA, Dewers TA, et al. Bacterial Diversity and Sulfur Cycling in a Mesophilic Sulfide-Rich Spring. *Appl Environ Microbiol* 2003; **69**: 5609–5621.
2. Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, et al. A new view of the tree of life. *Nat Microbiol* 2016; **1**: 16048.
3. Matheus Carnevali PB, Schulz F, Castelle CJ, Kantor RS, Shih PM, Sharon I, et al. Hydrogen-based metabolism as an ancestral trait in lineages sibling to the Cyanobacteria. *Nat Commun* 2019; **10**: 463.
4. Utami YD, Kuwahara H, Igai K, Murakami T, Sugaya K, Morikawa T, et al. Genome analyses of uncultured TG2/ZB3 bacteria in ‘Margulisbacteria’ specifically attached to ectosymbiotic spirochetes of protists in the termite gut. *ISME J* 2019; **13**: 455–467.
5. Oliver T, Sánchez-Baracaldo P, Larkum AW, Rutherford AW, Cardona T. Time-resolved comparative molecular evolution of oxygenic photosynthesis. *Biochimica et Biophysica Acta (BBA) - Bioenergetics* 2021; **1862**: 148400.
6. Needham DM, Yoshizawa S, Hosaka T, Poirier C, Choi CJ, Hehenberger E, et al. A distinct lineage of giant viruses brings a rhodopsin photosystem to unicellular marine predators. *Proceedings of the National Academy of Sciences* 2019; **116**: 20574 LP – 20583.
7. Bolger AM, Lohse M, Usadel B. Trimmomatic: a flexible trimmer for Illumina sequence data. *Bioinformatics* 2014; **30**: 2114–2120.
8. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 2012; **19**: 455–477.
9. Langmead B, Salzberg SL. Fast gapped-read alignment with Bowtie 2. *Nat Methods* 2012; **9**: 357–359.
10. Li H, Handsaker B, Wysoker A, Fennell T, Ruan J, Homer N, et al. The Sequence Alignment/Map format and SAMtools. *Bioinformatics* 2009; **25**: 2078–2079.
11. Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Res* 2015; **25**: 1043–1055.
12. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods* 2015; **12**: 59–60.

13. Pruesse E, Quast C, Knittel K, Fuchs BM, Ludwig W, Peplies J, et al. SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res* 2007; **35**: 7188–7196.
14. Eren AM, Kiefl E, Shaiber A, Veseli I, Miller SE, Schechter MS, et al. Community-led, integrated, reproducible multi-omics with anvi'o. *Nat Microbiol* 2021; **6**: 3–6.
15. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* 2020; **36**: 1925–1927.
16. Needham DM, Poirier C, Bachy C, George EE, Wilken S, Yung CCM, et al. The microbiome of a bacterivorous marine choanoflagellate contains a resource-demanding obligate bacterial associate. *Nat Microbiol* 2022; **7**: 1466–1479.
17. Wilken S, Yung CCM, Poirier C, Massana R, Jimenez V, Worden AZ. Choanoflagellates alongside diverse uncultured predatory protists consume the abundant open-ocean cyanobacterium *Prochlorococcus*. *Proceedings of the National Academy of Sciences* 2023; **120**: e2302388120.
18. Huerta-Cepas J, Szklarczyk D, Heller D, Hernández-Plaza A, Forslund SK, Cook H, et al. eggNOG 5.0: a hierarchical, functionally and phylogenetically annotated orthology resource based on 5090 organisms and 2502 viruses. *Nucleic Acids Res* 2019; **47**: D309–D314.
19. Gruber-Vodicka HR, Leisch N, Kleiner M, Hinzke T, Liebeke M, McFall-Ngai M, et al. Two intracellular and cell type-specific bacterial symbionts in the placozoan *Trichoplax* H2. *Nat Microbiol* 2019; **4**: 1465–1474.
20. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 2013; **30**: 772–780.
21. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; **25**: 1972–1973.
22. Whelan S, Irisarri I, Burki F. PREQUAL: detecting non-homologous characters in sets of unaligned homologous sequences. *Bioinformatics* 2018; **34**: 3929–3930.
23. Ali RH, Bogusz M, Whelan S. Identifying Clusters of High Confidence Homologies in Multiple Sequence Alignments. *Mol Biol Evol* 2019; **36**: 2340–2351.
24. Viklund J, Ettema TJG, Andersson SGE. Independent Genome Reduction and Phylogenetic Reclassification of the Oceanic SAR11 Clade. *Mol Biol Evol* 2012; **29**: 599–615.
25. Guindon S, Dufayard J-F, Lefort V, Anisimova M, Hordijk W, Gascuel O. New Algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Syst Biol* 2010; **59**: 307–321.

26. Kalyaanamoorthy S, Minh BQ, Wong TKF, von Haeseler A, Jermiin LS. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* 2017; **14**: 587–589.
27. Katoh K, Standley DM. A simple method to control over-alignment in the MAFFT multiple sequence alignment program. *Bioinformatics* 2016; **32**: 1933–1942.
28. Minh BQ, Schmidt HA, Chernomor O, Schrempf D, Woodhams MD, von Haeseler A, et al. IQ-TREE 2: new models and efficient methods for phylogenetic inference in the genomic era. *Mol Biol Evol* 2020; **37**: 1530–1534.
29. Wickham H. ggplot2: elegant graphics for data analysis. 2016. springer.
30. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017; **8**: 28–36.
31. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016; **13**: 581–583.
32. King WL, Siboni N, Kahlke T, Dove M, O'Connor W, Mahbub KR, et al. Regional and oyster microenvironmental scale heterogeneity in the Pacific oyster bacterial community. *FEMS Microbiol Ecol* 2020; **96**: fiae054.
33. Obiol A, López-Escardó D, Salomaki ED, Wiśniewska MM, Forn I, Sà E, et al. Gene expression dynamics of natural assemblages of heterotrophic flagellates during bacterivory. *Microbiome* 2023; **11**: 134.
34. Massana R, del Campo J, Sieracki ME, Audic S, Logares R. Exploring the uncultured microeukaryote majority in the oceans: reevaluation of ribogroups within stramenopiles. *ISME J* 2014; **8**: 854–866.
35. Logares R, Audic S, Santini S, Pernice MC, de Vargas C, Massana R. Diversity patterns and activity of uncultured marine heterotrophic flagellates unveiled with pyrosequencing. *ISME J* 2012; **6**: 1823–1833.
36. Rodríguez-Martínez R, Rocap G, Salazar G, Massana R. Biogeography of the uncultured marine picoeukaryote MAST-4: temperature-driven distribution patterns. *ISME J* 2013; **7**: 1531–1543.
37. Maliheh M, Ali AM, Rohit G, Abolhassan SFS, Francisco R-V. Genome Reconstruction from Metagenomic Data Sets Reveals Novel Microbes in the Brackish Waters of the Caspian Sea. *Appl Environ Microbiol* 2016; **82**: 1599–1612.
38. Blanco-Bercial L, Parsons R, Bolaños LM, Johnson R, Giovannoni SJ, Curry R. The protist community traces seasonality and mesoscale hydrographic features in the oligotrophic Sargasso Sea. *Front Mar Sci* 2022; **9**.



## Chapter 5

### **Long-term changes and interactions in the phytoplankton community at the Bermuda Atlantic Timeseries Study (BATS) between 1991 and 2020**

---

Fabian Wittmers<sup>1,2</sup>, Kevin L. Vergin<sup>9</sup>, Rachel J. Parsons<sup>7</sup>, Krista Longnecker<sup>8</sup>, Shuting Liu<sup>3,5</sup>, Jacqui Comstock<sup>3</sup>, Michelle L. Michelsen<sup>6</sup>, Luis M. Bolaños<sup>6</sup>, Elisa Halewood<sup>3</sup>, Jessica Godfrey<sup>7</sup>, Kevin Yongblat<sup>7</sup>, Ruth Curry<sup>7</sup>, Chang Jae Choi<sup>2</sup>, Nicolas Bates<sup>7</sup>, Rod Johnson<sup>7</sup>, Ben Temperton<sup>6</sup>, Hilary G. Close<sup>11</sup>, Leocadio Blanco-Bercial<sup>7,12</sup>, Elizabeth B. Kujawinski<sup>10</sup>, Craig A. Carlson<sup>3</sup>, Stephen J. Giovannoni<sup>4</sup>, Alexandra Z. Worden<sup>1</sup>

<sup>1</sup>Marine Biological Laboratory, Woods Hole, MA 02543, USA

<sup>2</sup>Ocean EcoSystems Biology Unit, RD3, GEOMAR Helmholtz Centre for Ocean Research Kiel, Kiel 24148, DE

<sup>3</sup>Marine Science Institute, Department of Ecology, Evolution and Marine Biology, University of California, Santa Barbara, CA 93106, US

<sup>4</sup>Department of Microbiology, Oregon State University, Corvallis, OR 97331, US

<sup>5</sup>Department of Environmental & Sustainability Sciences, Kean University, Union, NJ 07083, US

<sup>6</sup>School of Biosciences, University of Exeter, Exeter, Devon, EX4 4SB, UK

<sup>7</sup>Bermuda Institute of Ocean Sciences, St. George's 98C3+8G, Bermuda

<sup>8</sup>Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, US

<sup>9</sup>Microbial DNA Analytics, Phoenix, OR 97535, US

<sup>10</sup>Department of Marine Chemistry and Geochemistry, Woods Hole Oceanographic Institution, Woods Hole, MA 02543, US

<sup>11</sup>University of Miami Rosenstiel School of Marine and Atmospheric Science, Miami, FL 33149, US

<sup>12</sup>School of Ocean Futures, Arizona State University, Tempe, AZ, US

## Abstract

Long-term timeseries studies are essential for understanding changes in environmental conditions and microbial communities that occur as a consequence of natural and anthropogenic forcing. The Bermuda Atlantic Time-series Study (BATS) site where the BIOS-SCOPE project focuses on monthly sampling of physical, chemical, and biological parameters over depth profiles, has a molecular record sampled since 1991. We analyzed microbial taxa at BATS over the past 30 years using a newly sequenced 16S rRNA gene amplicon sequencing record, focusing on the V1V2 region, stretching from the surface to the mesopelagic ocean. Our analyses Capture cyanobacteria and plastids from eukaryotic phytoplankton, alongside non-photosynthetic bacteria, and thereby enables a comprehensive overview of the phytoplankton community and its interactions with bacteria over three decades. This revealed changes in the phytoplankton community in association with environmental parameters observed at BATS, including a phase of rapid warming between 2010 and 2020, and a subsequent decline in nutrient inventories in surface waters, from increased thermal stratification. Focusing on the summer surface, we observed a significant decrease in species richness of the total community and within the phytoplankton community specifically. This coincided decrease in cyanobacteria and eukaryotic phytoplankton relative abundance over time. *Prochlorococcus*, the major cyanobacterium during the stratified months, showed a decline in all of its major strains over time, with the emergence of a new, abundant strain during the period of accelerated warming not negating this overall decline. Within eukaryotic phytoplankton, putatively mixotrophic chrysophytes increased significantly since the onset of faster warming around 2010, with an increase especially within the uncultured marine clades A, B, D and E. This suggests that chrysophytes, alongside some mixotrophic members of the dictyochophytes and prymnesiophytes might be able to capitalize on the change in conditions by counteracting a decline in available nutrients through phagocytosis, while purely autotrophic lineages struggle. Using a co-occurrence network, we reveal interactions between phytoplankton and heterotrophic bacteria, with a general trend towards positive interactions between cyanobacteria and heterotrophs and negative interactions between eukaryotic phytoplankton and heterotrophs. Bacteria that were specific in their interactions with phytoplankton showed a significant change over time corresponding to the phytoplankton's trend, suggesting that the changes in the phytoplankton observed at BATS, coinciding with increased warming, have a downstream impact on the heterotrophic community.

## Introduction

Climate change, driven increasingly by greenhouse gas emission through anthropogenic activity, primarily the combustion of fossil fuels, has caused substantial warming of global surface temperatures averaging 1.1°C in 2011-2020, compared to the reference timeframe between 1850 and 1900 [1]. Biogeochemical and ocean ecosystem models predict a cascade of secondary effects due to warming of the surface ocean as a consequence of climate change [2, 3]. How these changes impact phytoplankton, forming the base of the marine foodweb, has become a major concern in the last decade [3–5]. Global observations indicate a decline in phytoplankton (based on chlorophyll measurements) co-occurring with an increase in surface ocean temperature [6, 7]. Warming impacts phytoplankton directly, depending on their temperature tolerance, and indirectly by increasing stratification of the water-column, which reduces the vertical input of nutrients through mixing [8], causing a decrease in nutrient inventory in the sunlit ocean [4]. These conditions generally decrease primary productivity with a more heterotrophic food web centred around recycling instead of export of organic matter [9]. This has major implications for the entire foodweb and cascading effects, for example impacting carbon potential fisheries yields globally [1].

Long-term timeseries studies are essential for understanding changes in environmental conditions and microbial communities that occur as a consequence of natural and anthropogenic forcing. They are key in both providing underlying data enabling large-scale modeling efforts and in verifying predictions made through these models on effects of climate change. The BIOS-SCOPE project focuses on bi-monthly sampling of physical, chemical, and biological parameters over depth profiles at the Bermuda Atlantic Time-Series Study (BATS) site. BATS is located at 31°40'N 64°10'W in the Sargasso Sea within the North Atlantic Ocean, providing an important timeseries for an open ocean ecosystem not influenced by larger landmasses. BATS, experiences seasonality throughout the year, including a deep mixing period each winter and a stratified water column during the summer months, with transition periods between, leading to considerable fluctuations in the microbial community throughout the year [10, 11]. This presents a contrast compared to the second long-term open ocean timeseries located in the Pacific Ocean. The Hawaii Ocean Timeseries Study (HOTS) site is located further south in the subtropical gyre, with only little seasonality [12]. HOTS is not subjected to the same annual deep mixing as observed at BATS, with more continuous stratification throughout the year and subsequently more oligotrophic conditions than at BATS. BATS shows dynamic seasonality with yearly fluctuations in the microbial community, including the composition of phytoplankton [10, 11]. During winter deep

mixing, the phytoplankton community is dominated by Class II prasinophytes, overtaken by dominant cyanobacteria, primarily *Prochlorococcus* and various relatively less abundant eukaryotic groups during the stratified summer season [11]. Nutrient concentrations reach their lowest concentrations in the surface ocean during the summer, but dissolved organic carbon accumulates at the same time, with extensive recycling of organic matter by the heterotrophic community [10]. With the onset of the autumn transition period, phytoplankton species that have persisted at low abundances in the deep chlorophyll maximum (DCM) over the summer seed the winter phytoplankton community benefiting of increase nutrient concentrations introduced vertically through deep mixing [10, 11].

This chapters focuses on characterising the surface microbial community at BATS over the past 30 years and its potential response to changes in abiotic factors, primarily temperature and secondary effects arising from it. The analysis will center around changes in phytoplankton including both cyanobacteria and eukaryotic phytoplankton lineages, as well as the heterotrophic community that interacts with photosynthetic lineages.

I will explore potential explanations for and consequences of the observed changes, specifically focusing on the importance of the photosynthetic community, primarily the surface during the stratified season for which the longest sampling-record exists. I hypothesize that phytoplankton lineages respond differently to changes in temperature, depending on their taxonomy, lifestyle and ability to compensate reduced availability of nutrients as a consequence of increase stratification.

## Material & Methods

### Oceanographic sampling

Samples were collected during research cruises aboard the R/V *Weatherbird I & II* and R/V *Atlantic Explorer* as part of the BATS timeseries project lead by the Bermuda Institute of Ocean Sciences (BIOS) and BIOS-SCOPE project. Conductivity-temperature-depth (CTD) profiling took place between Oct. 1991 and Aug. 2020 (Figure S1). Initially, DNA was collected at two depths (surface and 200m) until Aug. 2005, using McLane in-situ pumps and 142mm 0.2µm filters. From Nov. 2008 onwards, sampling was expanded to eight depth and twelve depths from Oct. 2016 onwards, using Niskin bottles mounted to a CTD rosette. During Niskin-bottle sampling, 4 L of water were filters through 0.22 µm Sterivex™ (Millipore) filters and frozen at -80°C until processing.

### Physical framework

Physical and biogeochemical data and CTD profiles were obtained from BATS (via <https://bats.bios.asu.edu/bats-data/>) and processed in R using tidyverse tools [13]. Structuring the water column into vertical zones, the surface mixed layer (ML) refers to depth from the surface to the base of the surface mixed layer where the density reached the surface density + 0.125 kg m<sup>-3</sup>. The deep chlorophyll maximum layer (DCM) comprises the part of the water column where the in-situ fluorescence reached at least 35% of the peak chlorophyll fluorescence, with the layer between the bottom of the ML and the top of the DCM being referred to as the upper stratified euphotic zone (USEZ). Below the DCM, the winter mode water layer (WMW) comprised layer between the base of the DCM and the density of the local deep winter mixed layer, 26.32 kg m<sup>-3</sup>. Seasonality was assigned to each cast within the timeseries based on CTD profiles, according to the following definitions: (1) Deep mixing begins when the mixed layer depth (MLD) goes deeper than the DCM layer depth and ends when the MLD abruptly shoals to around 100 m, resulting in the beginning of the (2) spring transition. The spring transition ends when the MLD shoals above the DCM for the remainder of the summer, initiating the (3) stratified season, which ends when the MLD gets deeper than the DCM. The (4) fall transition is defined as the period between the first entrainment of the DCM into the mixed layer until the last appearance of the DCM layer itself.

### DNA extraction and sequencing

DNA samples were extracted using a phenol chloroform-based protocol [14]. DNA was amplified using V1V2 16S rRNA gene primers 27F (AGAGTTTGATCCTGGCTCAG) and 338RPL (GCWGCCWCCCGTAGGWT) with ‘general’ Illumina adapters. Samples were pooled to equimolar concentrations and sequenced on the Illumina MiSeq (MiSeq Reagent Kit v2) platform. Raw sequencing reads were processed, trimmed (--p-trun-len-f 180 --p-trun-len-r 150) and amplicon sequence variants (ASVs) generated using the DADA2 R package (v1.14.0; [15]).

### Phylogenetic classification of amplicons

Quality controlled, representative V1-V2 16S rRNA gene ASVs were phylogenetically classified using a Python3 version (available through github: [github.com/BIOSCOPE/PhyloAssigner\\_python\\_UCSB](https://github.com/BIOSCOPE/PhyloAssigner_python_UCSB)) of PhyloAssigner [16], as published in [11], using a series of maximum-likelihood reference trees and alignments. To separate plastid & cyanobacterial ASVs from other bacterial ASVs, sequences were initially placed on a global 16S rRNA gene reference tree [16]. Cyanobacterial and plastid sequences were separated through a better-resolved tree composed of plastid and cyanobacterial 16S rRNA gene sequences [17]. Eukaryotic phytoplankton was then classified using a series of reference trees specific for chrysophytes, Viridiplantae (prasinophytes) [11], stramenopiles (and specifically for subsets within stramenopiles: dictyochophytes, pelagophytes) [17]. Cyanobacteria and specifically *Prochlorococcus* ecotypes were further classified through two consecutive reference databases [14, 18]. For chrysophytes, a new reference phylogeny was constructed, based on 41 chrysophyte 16S rRNA plastid sequences covering the V1V2 amplicon region, using two eustigmatophyte plastid sequences (FJ612231.1, HM127019.1) as an outgroup. Sequences were aligned using MAFFT (v7.310; [19]) and trimmed using trimAl (v1.4.1; [20]). A maximum-likelihood reference phylogeny was inferred using RAxML (v7.0.4; [21]) under the GTR model with 1000 bootstrap replicates based on 328 positions.

For chrysophytes and dictyochophytes, phylogenetically placed sequences were grafted on reference trees using gappa (v0.8.4; [22]) with the subcommand ‘examine graft’ and the best placement results, not resolving branches consisting of only placed sequences. The tree was further refined in R, removing all grafted ASVs not detected in at least 10 samples throughout the visualised data using the ‘drop.tip’ function of the ‘treeio’ R package (v1.23.0; [23]). All phylogenetic trees were visualised and amended with accessory data through R using the

‘ggtree’ (v3.6.1; [24]) and ‘ggtreeExtra’ (v3.19; [25]) packages. Genus-level taxonomy for prymnesiophytes is based on the best homology to a known reference species as determined by blasting all prymnesiophytes ASVs against NCBI’s nr database using blastn (v2.15.0 [26]).

### Diversity metrics

For surface water diversity calculations, 275 samples were rarefied to 25,000 reads per sample. Five samples were removed from the surface sub-dataset as they didn’t reach the required sequencing depth. All rarefaction curves of samples retained reached saturation, with a final rarefaction curve slope at 25,000 reads of 0.01 or below. Using the rarefied surface dataset, alpha diversity was calculated in R using the ‘diversity’ function in vegan (v2.6-2; [27]), computing both the Simpson, Inverse-Simpson, and Shannon-Wiener index. Species exchange rate (SER) was calculated as defined in Hillebrand et al. [28]).

$$SER = \frac{S_{imm} + S_{ext}}{S_{tot}}$$

where  $S_{imm}$  is the number of ASVs newly detected in the later sample,  $S_{ext}$  is the number of ASVs missing in the later sample and  $S_{tot}$  is the number of ASVs found in either one or both of the samples. A smaller SER indicates more species persisting, with a higher number indicating a higher turnover of species. Principle-coordinate analysis (PCoA) was conducted using a bray-curtis distance matrix computed through vegan’s ‘vegdist’ function and ordinated using the ‘cmdscale’ function. All diversity data and ordinations were visualised using ggplot2 (v3.5.1; [29]). Statistical significance for influence of environmental variables on ordinations were computed using non-parametric multivariate analysis of variance through ‘adonis’ (v2; [30]). General additive models and statistical comparisons between groups therein were computed using the ‘mgcv’ package (v1.8-40; [31]).

### Co-occurrence network analysis

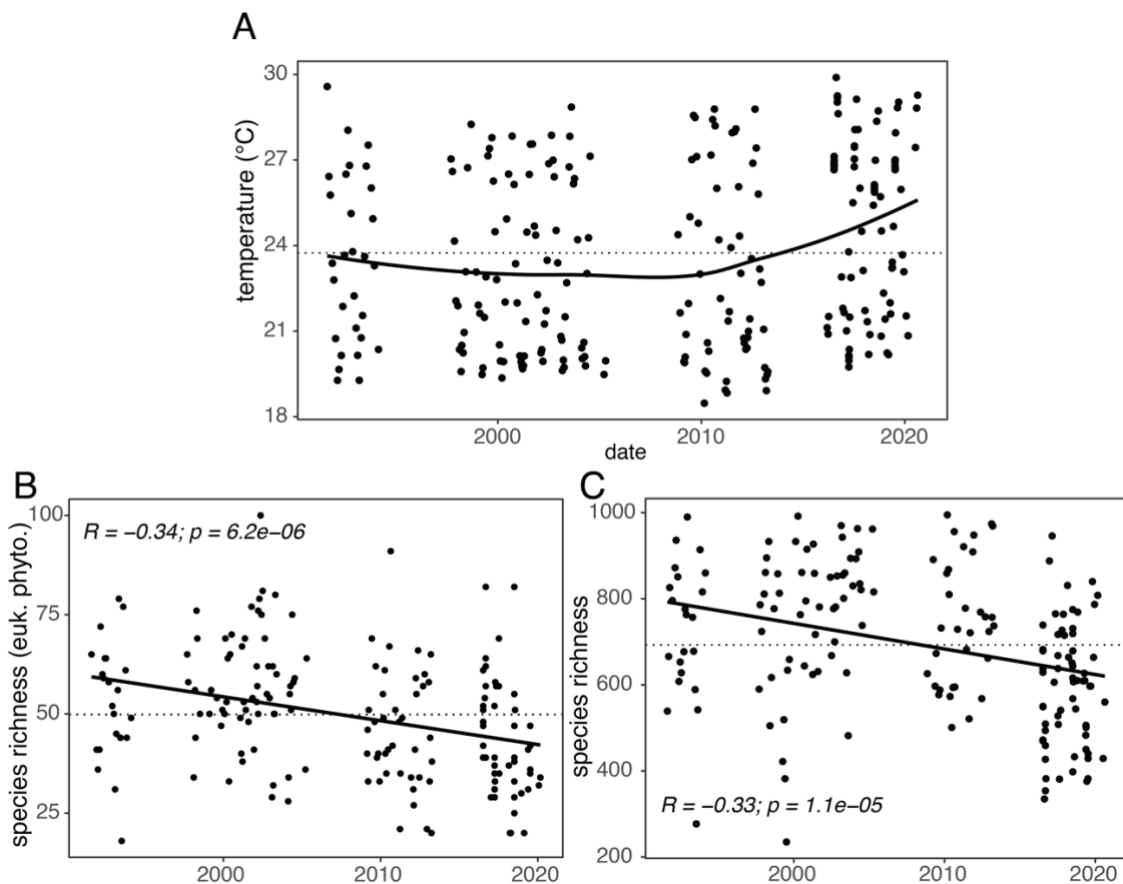
Co-occurrence networks were based on concepts as published by Faust et al. [32], with modifications specific to marine datasets as described in Lima-Mendez et al. [33]. The amplicon feature table was filtered to only include surface samples ( $\leq 5\text{m}$  depth), the only photic-zone layer sampled since the onset of the timeseries, resulting in 271 samples considered for network construction in total. Next, we separated the photosynthetic (both cyanobacteria and eukaryotic phytoplankton) and heterotrophic community into two separate abundance metrics, allowing targeted analysis of interactions between ASVs of different

lifestyles. The initial matrix of all phytoplankton ASVs comprised 5,443 ASVs, with the matrix of heterotrophic bacteria including 95,547 ASVs. ASVs were removed from the abundance matrix when they were not present in a minimum of 20% of samples, resulting in a total of 79 (phytoplankton) and 879 (heterotrophic bacteria) ASVs. Relative abundances were calculated afterwards, normalized to the total ASV read counts per sample in the respective matrix. For network computation, we opted against reconstruction time-delayed responses in biotic interactions, because we consider the duration of two weeks to 1 month between sampling timepoints too large to reliably resolve biotic responses, as argued in Deutschmann et al. [34]. Furthermore, years without sampling and uneven sampling intervals due to cruise schedules and weather influence would have required the extrapolation and simulation of large amounts of data, something we do not consider accurate in reflecting species interactions and community composition. Interactions between two ASVs were detected using Spearman's rank correlation. Significance was determined using permutation/renormalization followed by bootstrapping with 1000 iterations each. Taxon abundances were reordered and re-normalized to one after permutation. This generated null distributions and normalized abundances avoiding the detection of correlations solely based on the compositional structure of the input relative abundance data. Bootstrapping was used to detect 'real' significant correlations based on confidence intervals. *p*-values were corrected to adjust for false positive discovery rates using pooled-variance Z-test and interactions were considered significant with an adjusted *p*-value  $\leq 0.05$ . Interactions detected between two ASVs can be driven by both species responding in the same or similar way to a change in an environmental condition, resulting in so-called environmental triplicates (2 ASVs and 1 environmental factor). To remove these artefacts and identify only true biotic interactions between ASVs, we constructed a matrix for environmental factors, including nitrate  $\text{NO}_3^-$ , phosphate  $\text{PO}_4^{3-}$ , length of day (h of sunshine), salinity and temperature. Both matrices were compared against the environmental matrix in the same way as the comparison between phytoplankton and heterotrophic bacteria, and the 27,224 ASV-to-ASV interactions were subsequently filtered to remove all interactions in which both ASVs also correlated with the same environmental factor. This resulted in the removal of 24,863 interactions (91.3%) initially detected between two ASVs. It is likely that this also led to the removal of some interactions that are driven by a combination of overlapping environmental factors and biotic interactions that could not be resolved with the approach applied herein. Similarly, environmental triplicates can involve an environmental variable not considered in the

environmental matrix, leading to false positives (see discussion). All matrix-to-matrix comparisons were computed in R using ccrepe (v1.30.0; [35]) and final networks were visualised using igraph (v1.3.5; [36, 37]).

## Results & Discussion

Our analysis centred around the phytoplankton community and the bacteria that interact with it at BATS, and our analysis uses only samples from the surface ocean, which is the only sunlit depth-layer for which a molecular record exists since 1991 (Figure S1). BATS shows an increase in temperature at the surface waters (Figure 1A), with an acceleration in warming after 2010, a distinct change in the environmental conditions that has previously been noted [4]. This overall change in temperature and secondary effects over time could impact the microbial community through various mechanisms, with important implication for biogeochemical pathways impacted by microbial activity, including nutrient recycling, and carbon fixation, remineralization, and export.



**Figure 1: Change in temperature and species richness over time in the surface waters at BATS.** (A) change in temperature in analyzed surface samples (0-5m) at BATS. (B) Species richness of only eukaryotic phytoplankton and (C) the total community captured by V1V2 16 rRNA gene amplicon sequencing (heterotrophic bacteria, cyanobacteria, and eukaryotic phytoplankton). Dashed lines indicate the mean across all samples (dots) shown.

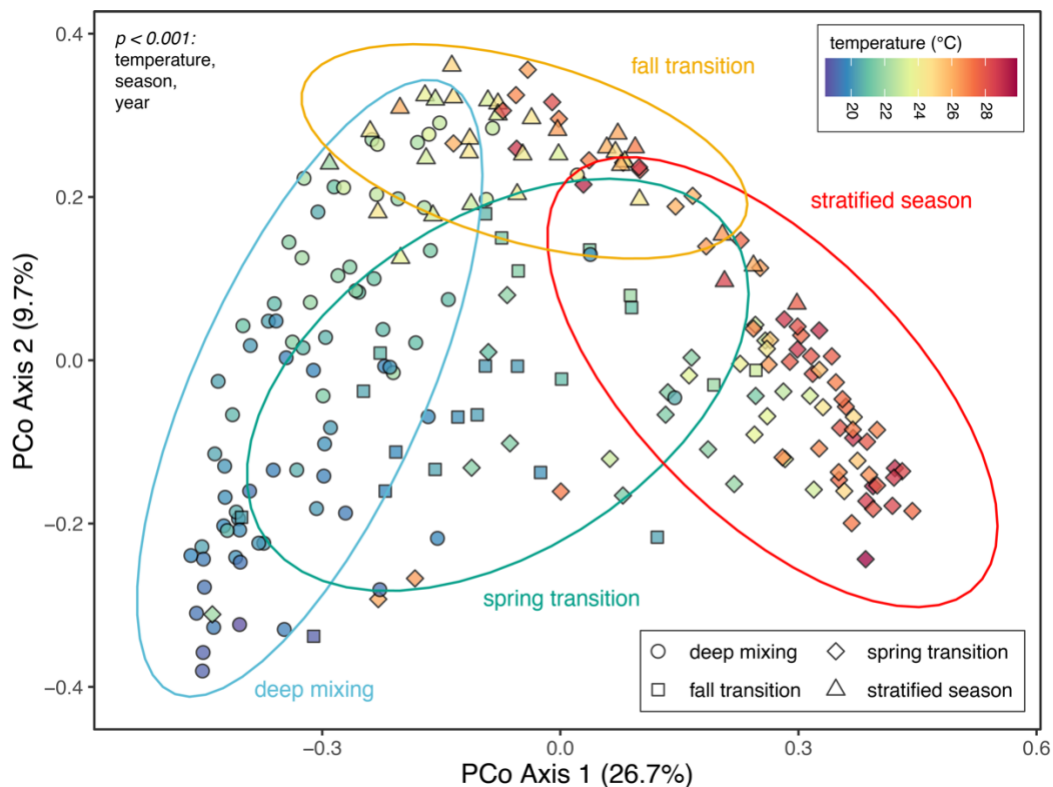
Focusing on the surface community for which we have the longest molecular record (starting in Oct. 1991), we initially investigated commonly applied general diversity metrics for alpha diversity (the diversity observed within a sample). We did not observe a significant change in any commonly applied diversity metrics (Shannon-Wiener-index [38], Simpson-index [39] or inverse Simpson index) over time when considering either the total or only the eukaryotic phytoplankton community. However, we did observe a significant decline in species richness (i.e., the number of species observed within a sample) over time in both eukaryotic phytoplankton ( $p < 0.001$ ) and the total microbial community ( $p < 0.001$ ) (Figure 1B, 1C). The Simpson diversity index is influenced disproportionately more by changes in evenness of common species than by diversity of rare species [28], suggesting that a decline in rare species is driving the decline in species richness (which essentially represents the number of observed species in a sample), explaining why we don't see a significant decline in any of the indices, but in species richness. The trend in species richness itself is likely biased, with species richness generally not capturing a decline over time because newly introduced species are detected earlier than species that are going extinct [40]. This suggests that the decline in species richness is likely an underestimation of the actual decline in richness taking place. Species richness itself also misses neutral species turnover, essentially the replacement of one species by another over time, which would not affect the number of total species observed at any given time. While we do observe a decline in species richness at the surface over time, however, this could be congruent with the disappearance of more species than detected and more change in community composition, i.e., neutral turnover, than detected by species richness.

#### Species turnover and change in community composition over time

We next attempted to explore turnover through species exchange rate (SER) [28] over time but did not observe a significant trend (Figure S2A). The SER is very susceptible to uneven sequencing depth, which we addressed through rarefying (randomized subsampling of samples with higher sequencing depth), but especially rare species are often not detected or lost through this process, inflating the SER [28]. Seasonality induces additional variability, with the SER lower during months within the stratification period versus more dynamic transition periods (Figure S2B). The molecular timeseries at BATS is not sampled at a strict frequency, with sampling intervals varying between 10 and 50 days throughout, due to weather and logistical constraints. Subsequently, with an increase in the gap between two

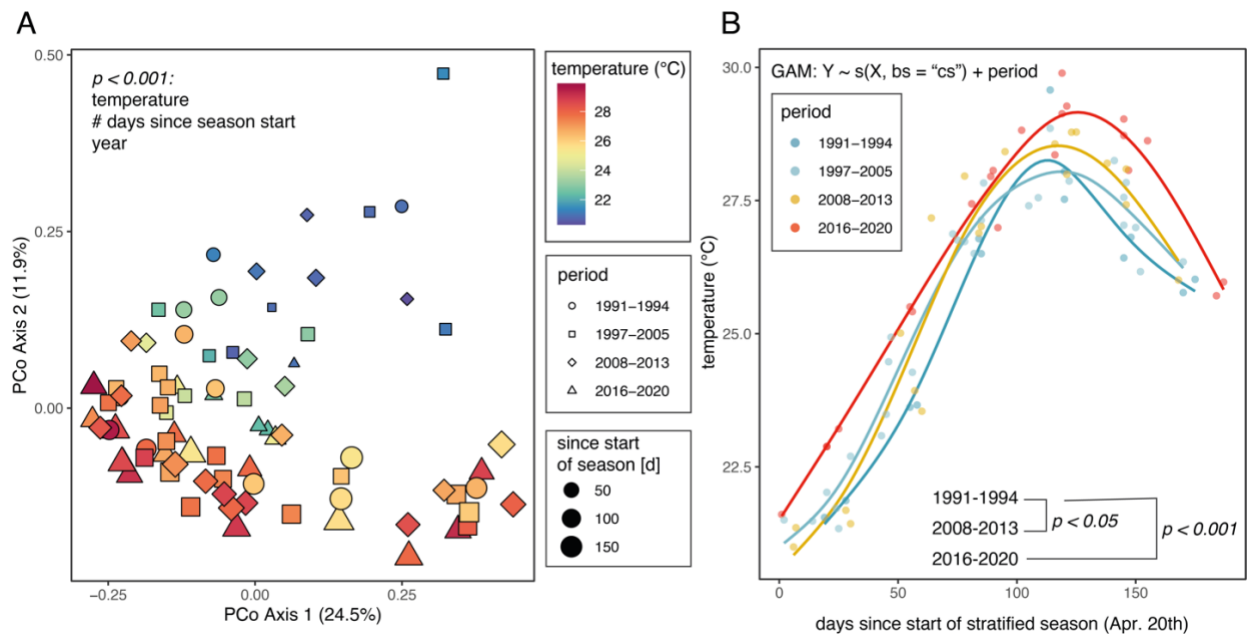
points considered to compute SER, we observed a significant ( $p < 0.01$ ) increase in SER between those timepoints (Figure S2C), skewing potential results in SER observations. Variability in sampling gap, seasonality at BATS and detection-limits for rare species bias and increase the variability of SER, blurring any true change in species exchange rates potentially present in the microbial community.

BATS, although located at a relatively low latitude (31°40'N 64°10'W), experiences substantial changes through seasonality, with deep mixing occurring during the winter months and strong stratification during the summer, with shorter spring and autumn transition periods in between [10]. This seasonality leads to major systematic change in the microbial community composition throughout each year (Figure 2), significantly driven by the fluctuations in temperature occurring simultaneously. The impact of seasonality, however, blurs potential trends occurring over longer time periods, with yearly variability being a significant factor driving differences in the community composition.



**Figure 2: Differences in surface community composition throughout seasons.** Principal coordinate analysis (PCoA) of 270 rarefied surface samples (25,000 reads) collected between 1991 and 2020 between 0 and 5 m. Exact sampling depth and the exact time of sampling were not significant in permutation testing of PCoA positions. The impact of nutrient concentrations was not tested due to concentrations below the detection limit for surface samples, especially during the stratified (summer) season.

To focus our investigation on possible change over time, we separated the timeseries dataset by season, mainly relying on the stratified summer season, for which the most consistent sampling recorder exists, including designated process cruises throughout recent years. Principle coordinate analysis (PCoA) of surface samples recovered during the stratified period, the longest seasons at BATS, stretching generally from the middle of April until the end of October (up to 195 days) revealed significant changes in community composition during the season (Figure 3A). Throughout the stratified season, the surface water warms up continuously. This increase in temperature is a significant driver of changes observed in the community composition throughout the season. We observed significant variability on a year-to-year basis, with some indications that the community varies in how it changes throughout the season from year to year, while the initial community at the beginning of the season is more similar than at the end. Although physically a stability period, this does not seem to translate to the microbial community which seems to stay dynamic throughout the summer. Inorganic nutrient availability ( $\text{NO}_3^{2-}$ ,  $\text{PO}_4^{3-}$ ,  $\text{Fe}_3^+$ ) is likely another driver for this change, but concentrations are generally below detection limit so this could not be tested statistically. It is to be expected that the nutrient inventory in the surface continuous to decline from already low starting levels at the beginning of the stratified season, further driving competition for these scarce resources, which likely impacts the microbial community. Notably, the temperature during the stratified season was significantly higher in 2008-2013 and more so, the most recent sampling period (2016-2020), compared to temperatures measured in the initial sampling period 1991-1994 (Figure 3B).



**Figure 3: surface community changes during the stratified stability period.** (A) Principal Coordinate Analysis of all samples collected during the stratified period between 1991 and 2020. Samples were grouped into sampling periods and are colored based on temperature at sampling and sized based on when sampling occurred relative to the start of the stratified period. Significance was tested using a permutation test [30]. Exact sampling depth and time of sampling were not significant factors. Nutrient concentrations were not tested because they did generally not reach detection limit during the summer season. (B) Relationship between temperature in the sampled water and the time since the start of the stratified period. Curves were fitted using a general additive model (GAM), distinguishing between the four sampling periods. Statistical significance was tested using a permutation test on the fitted lines, determining that the period 2008-2013 ( $p < 0.05$ ) and the latest sampling period 2016-2020 ( $p < 0.001$ ) have a statistically significantly different parametric coefficient from the first sampling period (1991-1994).

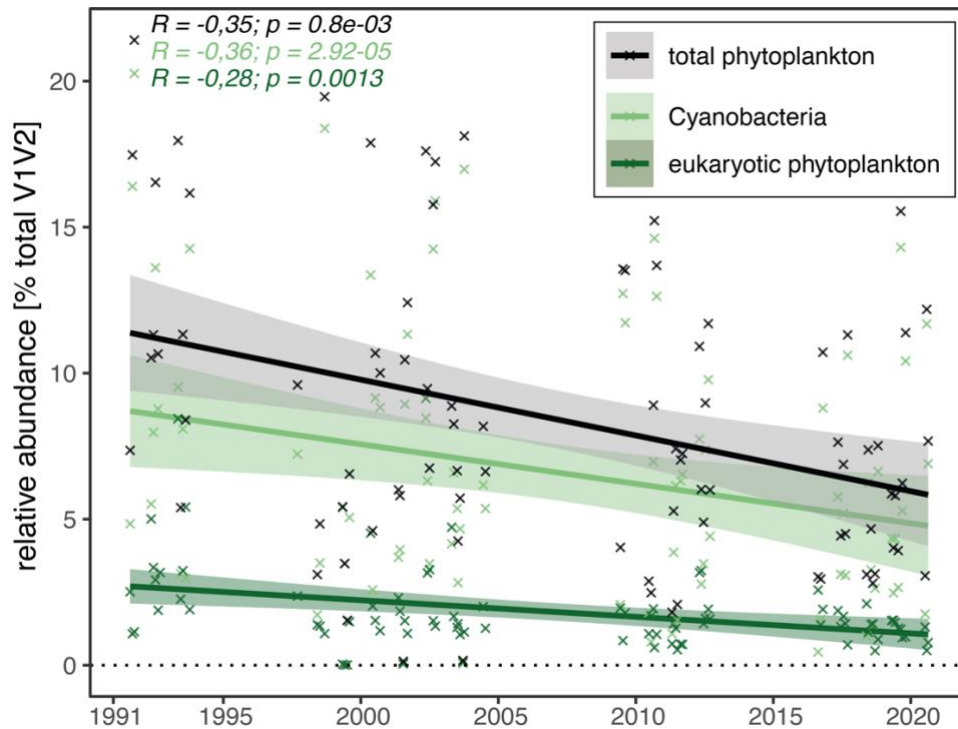
The analysis here on out will focus on the phytoplankton community specifically, predicted to be particularly heavily affected by climate change [3, 5]. We aimed to investigate what effects the changes in biogeochemical parameters, particularly the increased warming during summer and connected secondary effects observed at BATS have on the phytoplankton community that forms the base of the microbial foodweb. The second part of the following analysis will center around the interactions of phytoplankton with heterotrophic bacteria, to investigate potential downstream effects on the community composition driven by changes in phytoplankton.

### Relative decline in surface phytoplankton

The open ocean phytoplankton community in lower latitudes is generally dominated by tiny cyanobacteria, particularly *Prochlorococcus* lineages (so-called ecotypes) [41], reaching their

highest abundances during the summer stratified periods. Although *Prochlorococcus* is most abundant in absolute cell numbers between 50m and 100m BATS, depending on the ecotype [42], it contributes a similarly relative portion of phytoplankton in the stratified season at the surface (Figure S3A), where it is well-adapted to challenging environmental conditions, compared to (larger) eukaryotic phytoplankton. During the stratified period, nutrient concentrations are depleted below detection limit while there is no upwelling or other vertical transport of nutrient rich water from below, with *Prochlorococcus* outcompeting other marine cyanobacteria like *Synechococcus* and eukaryotic phytoplankton, due to their large surface-to-volume ratio (i.e., small cell size,  $\sim 0.8\mu\text{m}$ ) [43]. The surface *Prochlorococcus* population itself is dominated by members of the ecotype high-light II (HLII) (Figure S3B), which is best adapted to the challenging environmental conditions, including extremely low nutrient conditions, high temperatures and high light intensity and UV radiation [44–46]. At 40m and 80m, HLI is another high-light adapted ecotype increasing in relative abundance relative to the surface. At 80m and increasing below that, multiple low-light adapted ecotypes, including LLIII, LLI and especially LLVII (primarily subclade A), are the main ecotypes, with high-light ecotypes still present, but making up only 10-20% of the *Prochlorococcus* population at 160m. Cells of low-light *Prochlorococcus* ecotypes are optimized for growth at low-light intensities, far below the ideal light levels of high-light ecotypes. Their cells are generally larger and contain more chlorophyll pigments for optimized light-harvesting [47].

Between 1991 and 2020, we observed a decline in the relative abundance of phytoplankton (cyanobacteria and eukaryotic phytoplankton) compared to heterotrophic bacteria at the surface (Figure 4), coinciding with the warming of surface waters observed (Figure 1A). Both, a decline in cyanobacteria (almost exclusively *Prochlorococcus*) and a decline in eukaryotic phytoplankton contribute to this overall decline. We observed the lowest relative abundances in phytoplankton in the last decade, coinciding with an acceleration in warming observed at BATS since 2010 [4]. We did not observe a shift towards cyanobacteria relative to eukaryotic phytoplankton previously described for the integrated water column [4], but rather a shift towards to a more heterotrophic community in the surface overall, with all groups of phytoplankton seemingly struggling with the change in environmental conditions, in line with general concepts in how marine ecosystems will respond to warming [9].



**Figure 4: Long-term changes in relative phytoplankton abundance at BATS surface during the stratified season.** Relative abundance of total phytoplankton (cyanobacteria and eukaryotic phytoplankton combined), and cyanobacteria and eukaryotic phytoplankton separately, out of all V1V2 16S rRNA over time.  $p$ -values are based on linear correlations. Shaded areas indicate 95% confidence intervals. Points indicate relative abundances for each day sampled during the stratified season. When multiple samples were taken on the same day, e.g., during process cruises at BATS between 2016 and 2020 as part of BIOS-SCOPE, values were averaged.

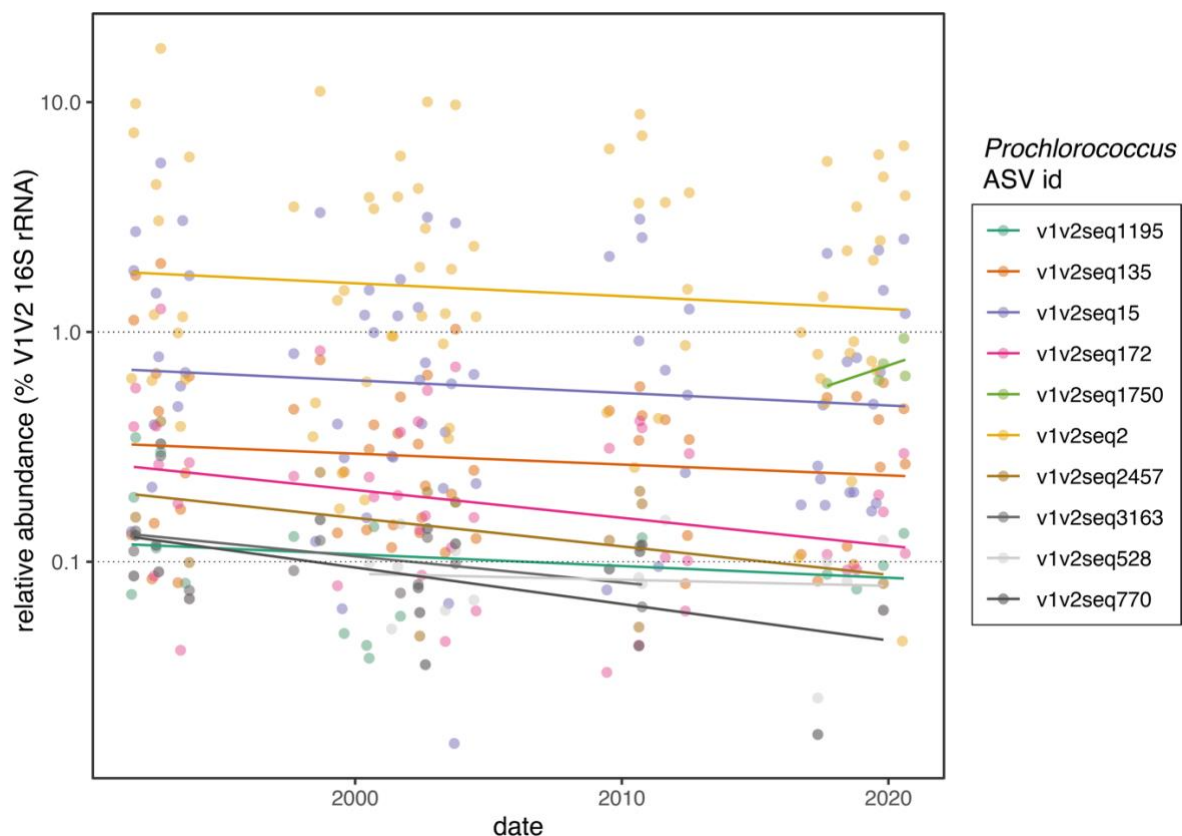
We next focused on a more in-depth taxonomic breakdown of both the cyanobacteria and eukaryotic phytoplankton lineages present. The *Prochlorococcus* population at BATS consist of multiple distinct ecotypes, and similarly, the eukaryotic phytoplankton is composed of various taxonomic groups with different lifestyles, environmental niches, and optimal environmental conditions, suggesting that the change observed over time likely did not affect all strains of phytoplankton the same. Similarly, some lineages could even be benefiting of the reduction or even disappearance of competing strains and therefore increase in relative abundance.

#### Changes in *Prochlorococcus* ecotype composition

During summer and at the surface, all abundant strains within the *Prochlorococcus* population belong to the ecotype HLII. One strain (i.e., one ASV) was particularly abundant (v1v2seq2), but multiple other ASVs were major contributors to the overall *Prochlorococcus*

population. The overall decline in relative abundance of *Prochlorococcus* in the summer surface waters we observed seems to be driven by a decline major *Prochlorococcus* ASVs (significant decline  $p < 0.001$  for v1v2seq2 and v1v2seq15), and not by a decrease in v1v2seq2 alone (Figure 5). ASV v1v2seq1750 is a noteworthy exception to this trend, not present in the dataset before 2011 at any depth, but then appearing and becoming the third most abundant surface *Prochlorococcus* (within HLII and generally) ASV by the end of the timeseries in 2020. Although increasing in abundance over time, the occurrence of this new strain did not negate the decline in all other strains, still resulting in a net decline in *Prochlorococcus*. v1v2seq1750 is not a rare species (approaching a mean relative abundance of 1% total V1V2 rRNA in the summer of 2020), so its presence only in the later part of the timeseries cannot be explained by issues in detection of rare species, suggesting that this is a true new abundant strain previously not present at BATS. Theoretically, the strain could have been present the whole time or at least for some period before increasing significantly in abundance but remained at extremely low (below detection limit) concentrations for decades, until it became an important strain. More likely, this represents the evolution or introduction of a new strain into the *Prochlorococcus* population, perhaps due to a change in environmental conditions favoring this strain vs strains that are unable to adapt to the warmer, more stratified conditions at BATS.

The occurrence of new sub-species variants could be explained by evolution, or horizontal introduction of the variant from other environments, e.g., through the movement of water masses through currents or eddies, which transport a distinct water mass in its center and can introduce new species this way [48]. This theory is particularly appealing for eddies originating south of BATS and transporting a community of microbes adapted to warmer conditions north. BATS is frequently impacted by eddies and horizontal movement of water, making this a plausible vector of introduction [49], but would require the strain to be established in another environment to be exported from. In an attempt to establish if v1v2seq1750 has been observed in environmental surveys before, we queried extensive V1V2 16S rRNA reference datasets from the North Atlantic, the Caribbean Sea [50], the North Pacific [51], Indian Ocean [18] and Mediterranean [52] but were unable to confirm the presence of the ASV in any of them. We also didn't obtain any exact match to reference sequences in the NCBI nr database. Although it cannot be considered a proof, this opens the possibility that this sequence might reflect an evolutionary response to the changing environmental conditions. New, including some abundant *Prochlorococcus* strains are still being discovered, leaving this question open for more comprehensive analysis in the future.



**Figure 5: Changes in *Prochlorococcus* ASVs over time.** Relative abundance of the top ten *Prochlorococcus* ASVs between 1991 and 2020. Ecotype classifications are based on PhyloAssigner [11]. v1v2seq2 and v1v2seq15 significantly declined over time ( $p < 0.001$ ).

### Changes in eukaryotic surface phytoplankton - the rise of the chrysophytes

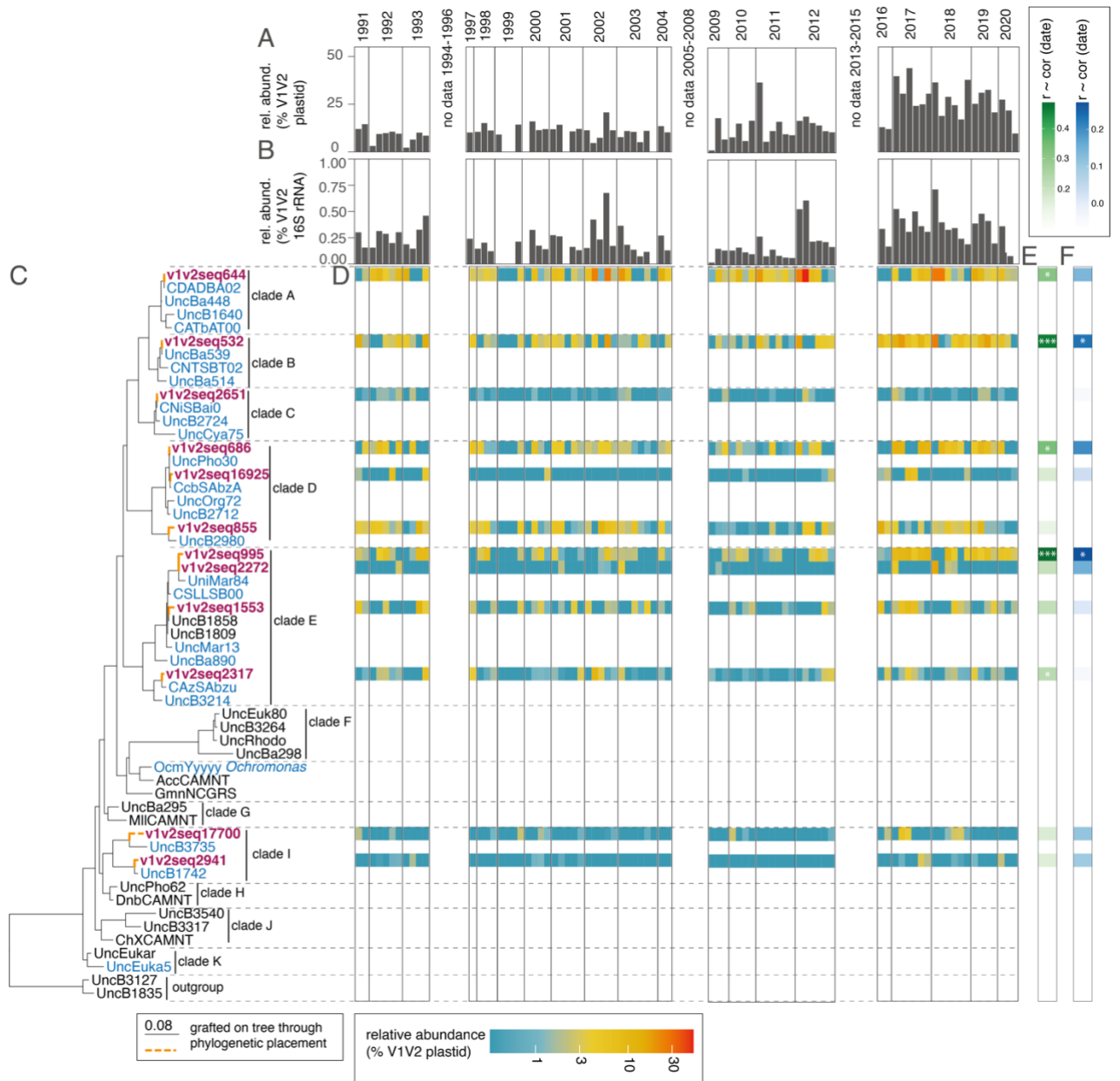
Even though eukaryotic phytoplankton is less abundant relative to cyanobacteria at the surface during the they are still an important component of surface phytoplankton.

Furthermore, eukaryotic phytoplankton cells are larger than cyanobacteria, subsequently making up disproportionately more organic material. As a consequence of their size, they also generate a larger phycosphere than smaller cells [53], opening more opportunities for interaction with other microbes, i.e., having a bigger microbiome. We observed changes within the eukaryotic phytoplankton community composition over time. Although the relative abundance of the eukaryotic phytoplankton decreases over time overall, this revealed some eukaryotic lineages that seem to be able to capitalize on the changing environmental conditions observed at BATS, subsequently rising in relative abundance over time (relative to the total photosynthetic community).

We observed a relative increase in chrysophytes in the surface eukaryotic community (Figure 6A, 6B). Chrysophytes are a class (Chrysophyceae) within the stramenopiles division

[54] mostly characterized as ‘golden algae’ in freshwater and brackish environments, but it also includes marine representatives that are relevant members of plankton community in ocean environments across all latitudes [54–56]. Marine chrysophyte diversity is poorly resolved, especially when relying on the 16S (plastid) rRNA gene as a marker. To enable high-level taxonomic resolution, we reconstructed a phylogenetic reference tree, relying mostly on environmental sequences of uncultured chrysophytes, followed by phylogenetic placement of putative chrysophyte ASVs from BATS using a rewritten and customized version of the PhyloAssigner tool [16], allowing (sub)-species variant resolution of detected chrysophytes (Figure 6C).

An overall increase in relative abundance over time occurred for most ASVs compared to previous sampling periods but was significant for representatives of chrysophyte clade B (v1v2seq532), clade E (v1v2seq995, v1v2seq2317), clade D (v1v2seq686), clade A (v1v2seq644) (Figure 6D). None of those clades comprise plastid sequences of cultured or otherwise known chrysophytes and gaining insights on their lifestyle is therefore challenging. Generally, Chrysophytes have larger cell size (genus *Ochromonas*: 2-30  $\mu\text{m}$  [57]), compared to lineages like *Prochlorococcus* ( $\sim 0.8\mu\text{m}$ ) [43] and they would subsequently at a disadvantage when competing for the extremely limited inorganic nutrients available in the stratified surface layer. In the context of increased warming and subsequently stronger stratification through reduced vertical mixing of water, which resulted in the depletion of the nutrient inventory at BATS [4], a decrease in chrysophytes, as in other phytoplankton, would be a logical consequence. The increase in abundance observed herein contradicts these predictions, with a mixotrophic lifestyle being one possible explanation for the trends observed.



**Figure 6: Chrysophytes become more important in the stratified summer surface ocean over time.** Relative abundance of all chrysophytes out of V1V2 16S rRNA plastid reads (eukaryotic phytoplankton) (A) and all V1V2 16S rRNA reads (B). (C) Maximum-likelihood phylogenetic reconstruction of chrysophytes based on 16S rRNA gene sequences. The phylogeny was reconstructed under the GTR model with 328 positions covering the V1V2 region. Sequences of marine origin are labelled in blue. ASVs recovered herein and grafted onto the phylogeny based on best node phylogenetic placement are highlighted in bold text and purple. (D) Heatmap showing the relative abundance (out of eukaryotic phytoplankton) for each separate ASV per sample. Correlation coefficient for correlation between the relative plastid abundance (E) and relative abundance (total) (F) of each ASV with the date of sampling. Significant correlations are marked with via asterisks (\*  $p \leq 0.05$ ; \*\*  $p \leq 0.005$ ; \*\*\*  $p \leq 0.0005$ ). Years that included designated processing cruises during the summer stratified period were subsampled to reduce the number of samples. Only one sample per cruise was considered.

Although the study of marine chrysophytes is difficult because of the lack of characterized isolates, some members have been identified as predatory mixotrophs [58, 59]. Mixotrophy is herein constraint to predatory mixotrophy, i.e., the ability of a protist to conduct photosynthesis through a plastid and phagocytise bacterioplankton (predation) [60, 61]. This enables larger cells to harvest light-energy by simultaneously acquiring inorganic nutrients (and potentially some essential organic nutrients like vitamins) through predation of bacterioplankton [62]. Field-based experiments confirmed that almost all (96%) of detected chrysophytes were able to feed on *Prochlorococcus* in the Pacific Ocean [55] and chrysophytes have been implied as actively feeding mixotrophs in both the Atlantic and Pacific subtropical oligotrophic gyres [58, 62]. Some chrysophytes have lost their ability to conduct photosynthesis [63], but our classification method, relying on 16S rRNA plastid sequences implies that the chrysophytes discussed here retained their plastids (while we do not detect potential heterotrophic chrysophytes here). While our data does not provide evidence for their potentially mixotrophic lifestyle, a previously characterized marine chrysophyte was found to be an obligate mixotroph, adapted to resource efficient growth in oligotrophic ocean environments [64]. The unknown environmental clades detected herein, especially clades B, D and E which have increase in relative abundance over time, probably utilize predatory mixotrophy, feeding on *Prochlorococcus* or other bacterioplankton to offset the competitive disadvantage because of their smaller surface-to-volume ratio.

A rise in putatively mixotrophic lineages like chrysophytes at BATS has implications for dynamics across the food chain, the transfer of carbon through the microbial foodweb, and subsequently marine carbon cycling in general. Mixotrophs as important contributors of bacterivory, and in the oligotrophic gyre in the North Atlantic, they are more important in controlling bacterivory than their aplastic counterparts because they are numerically more abundant [62]. An increase in the abundance of some mixotrophic lineages could indicate more predation on bacterioplankton like *Prochlorococcus*. Mixotrophs are capable of balancing loss of carbon during respiration of prey through photosynthesis and simultaneously supplement photosynthesis-limiting nutrients through grazing [65]. This leads to a higher carbon transfer efficiency through the food-chain, which, combined with the, on average, larger cell size of mixotrophs results in an increase in carbon export in model estimates [66]. Theoretically, this could counteract the reduced carbon export resulting from more organic matter recycling and less primary productivity expected for these conditions [9]. This could be an explanation for observations at BATS where there is no significant change in carbon export between 1991 and 2020 [4]. This hypothesis would require a depth-

integrated investigation of a potential increase in mixotrophs and a better understanding of actual cell sizes of lineages present. Cell-sizes within the chrysophyte genus *Ochromonas* range from 2-30  $\mu\text{m}$  [57], and those of environmental clades A – E and I present at BATS (Figure 6) are unknown because they only consist of uncultured lineages, making this a source of considerable variability that needs to be considered.

#### Dynamics in other putatively mixotrophic lineages

Chrysophytes are not the only putatively mixotrophic lineage present at BATS. An increase in mixotrophic chrysophytes would not automatically result in an increase in bacterivory if it comes at the cost of another mixotrophic lineage. To examine if mixotrophic lineage generally increased or decreased in relative abundance during the summer stratified season independent of their taxonomy, or if chrysophytes demonstrate an exemption, we investigated changes in other lineages that contain putative mixotrophs. Predatory mixotrophs have been described in various lineages, and in oligotrophic marine ecosystems mostly originate from the Bolidophyceae, Chrysophyceae, Dictyochophyceae, Prymnesiophyceae [58].

Additionally, Rhizaria and dinoflagellates are major lineages including mixotrophs present at BATS [67], but their abundances are challenging to assess through molecular approaches as used herein. Both lineages utilise diverse strategies to acquire plastids, and extremely poor reference data exists, making identification of 16S V1V2 plastid reads from either lineage challenging. E.g., plastids are frequently ‘stolen’ from other photosynthetic lineages, through a process called kleptoplasty [68]. Alternatively, 18S rRNA amplicon sequencing could be used to identify Rhizaria or dinoflagellates, but both generally encode extremely variable copies of the 18S rRNA gene (sometimes more than 1000 copies per genome [69]), making 18S rRNA as a quantitative tool essentially useless, and a comparison of their abundance vs lineages detected and relatively quantified herein not possible.

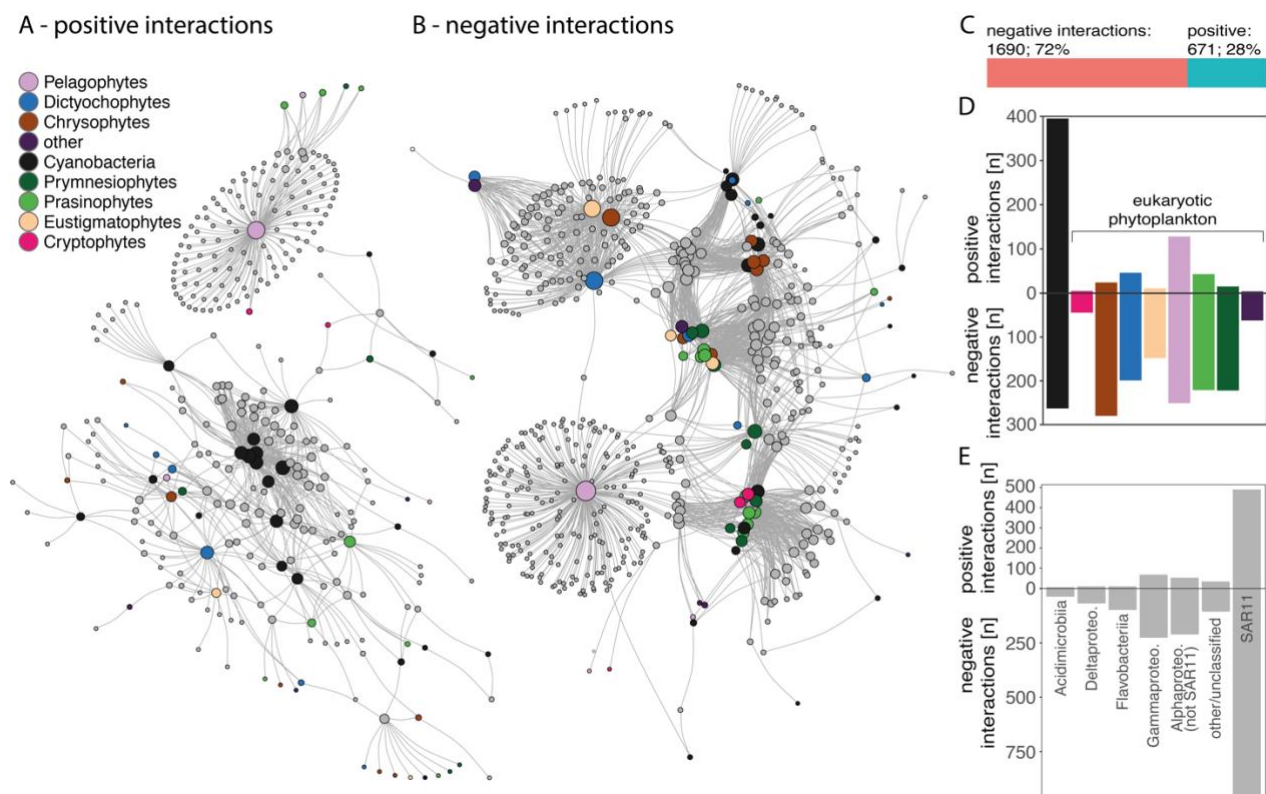
We were able to detect bolidophytes throughout the timeseries, but all representatives are rare species not continuously present throughout the timeseries, with an average relative abundance of  $0.70 \pm 0.82\%$  V1V2 plastid 16S rRNA for all Bolidophyceae ASVs in the surface. They were not investigated further herein because derived trends would be heavily influenced by biases induced by detection limits. Prymnesiophytes and dictyochophytes are the major eukaryotic groups present in the surface layer at BATS, apart from chrysophytes, with average relative abundances of  $20.17 \pm 8.74\%$  and  $26.93 \pm 16.01\%$ , respectively. On a cumulative level, neither dictyochophytes nor prymnesiophytes clades displayed a decline or

increase over time (Figure S4A, Figure S5A). Prymnesiophytes comprise multiple clades, including coccolithophores such as the model species *Emiliana huxleyi*, which are strictly autotrophic, and the *Phaeocystis* and *Chrysochromulina* clades which include mixotrophic members, the latter being one particularly diverse at BATS (Figure S4B) [70, 71]. Interestingly, we observed an increase in the relative abundance of all *Phaeocystis* ASVs we detected, while coccolithophores (*Emiliana huxleyi*) decreased (Figure S4C). Within *Chrysochromulina*, we observed more variability, with about half of the ASV, including the most abundant *Chrysochromulina* ASV (v1v2seq223) increasing significantly, while the rest declined in relative abundance. Generally, all clades without mixotrophic characteristics seemed to decline, while mixotrophic lineages either increased (*Phaeocystis*) or showed more diverse internal trends (*Chrysochromulina*). Within the dictyochophytes, we observed a decline in all clades but DEC II and *Rhizochromulina* (Figure S5B, S5C). DEC (Dictyochophyceae Environmental Clade) II represent one of the many dictyochophyte clades consisting only of environmental sequences and no isolates with known lifestyles, while *Rhizochromulina* are mixotrophs [58]. DEC II comprises the two main dictyochophytes (v1v2seq261, v1v2seq948) present in the summer surface at BATS, which, together with their increase in abundance could indicate a mixotrophic lifestyle. We observed the most significant decline in v1v2seq2894, part of one of two clades (Novel BATS clade I & II) that have no isolates or environmental reference sequences, suggesting that these might indicate novel, previously unknown clades.

Overall, observed trends for three lineages with known mixotrophic lineages (dictyochophytes, prymnesiophytes and, most notably, chrysophytes) do raise questions about the potential increase in mixotrophic protists during the summer in the BATS surface layer. Trends observed within dictyochophytes and prymnesiophytes, however, do show that we do not see an increase in all mixotrophic lineages across taxonomic groups, e.g., *Chrysochromulina*. While results herein suggest an overall increase, not all mixotrophic lineages seem to benefit similarly from the changes at BATS over the past decades, potentially reaching the limit of their temperature tolerance. The lack of known isolates for most of these clades of important mixotrophs makes predictions about the reason for and impact of changes in mixotrophic taxa challenging, highlighting the need for a deeper investigation and characterization of more, so far, purely environmental, clades.

### Co-occurrence networks reveal interactions between surface phytoplankton and bacteria

Next, we aimed to explore how surface phytoplankton at BATS interacts with heterotrophic bacteria and explore potential downstream effects of changes in abundance of some phytoplankton lineages. We reconstructed a co-occurrence network for interactions between heterotrophic bacteria and the photosynthetic community (cyanobacteria and eukaryotic phytoplankton) (Figure 7). Overall, this resulted in the detection of significant interactions between 72 phytoplankton ASVs and 567 heterotrophic bacterial ASVs, revealing both positive (Figure 7A) and negative (Figure 7B) associations, although negative interactions clearly dominated overall (Figure 7C). Interactions between two entities in co-occurrence networks can represent physical associations, such as the attachment of one cell to another, but are not restricted to those. The detection of more diverse interactions represents an advantage of co-occurrence networks as supposed to other approaches, such as single-cell genomics (see chapters 2, 3 and 4). Positive interactions in co-occurrence networks can comprise beneficial interactions between heterotrophic bacteria and phytoplankton, such as mutualistic symbiosis. Negative interactions can represent more diverse interactions, most prominently predation, although herein, focusing on phytoplankton, this would require a mixotrophic lifestyle for the phytoplankton organism involved in the interactions. Heterotrophic bacteria can drive negative interactions through pathogenicity and parasitism, which in the context of their interactions with phytoplankton, encompasses algicidal bacteria, potentially associated with the phycosphere of the phytoplankton cell. Competition for other limiting resources can be another driver of negative interactions, especially when the limiting resource, e.g., iron, are not considered in the filtration step to detect ‘environmental triplicates’. Interestingly, positive interactions were dominated by interactions with cyanobacteria (58.8%), and negative interactions comprised interactions between heterotrophic bacteria and cyanobacteria, but also a more diverse pool of various lineages of eukaryotic phytoplankton, including mainly chrysophytes, pelagophytes, prymnesiophytes, prasinophytes and eustigmatophytes (Figure 7D). The ASV with the most interactions total, v1v2seq57, was classified as *Pelagomonas* (Pelagophyceae), a presumed autotroph. This lineage is generally found at deeper depth during the summer and only becomes abundant in the deep mixing at the surface when deep ventilation introduces nutrients into the surface layer [17]. Because this analysis was focused on the main phytoplankton groups observed during the stratified summer season with very limited nutrient supply, it was not investigated further.



**Figure 7: interactions between phytoplankton and bacteria in the surface ocean at BATS.** (A) Positive and (B) negative interactions detected between ASVs of heterotrophic bacteria and photosynthetic lineages (cyanobacteria and eukaryotic phytoplankton), not considering interactions paralleled by environmental co-correlation of the interacting species. Nodes are scaled based on the number of interactions. (C) proportion of negative and positive interactions observed. Taxonomic breakdown of the positive and negative interactions from the (D) phototrophic, or (E) heterotrophic side.

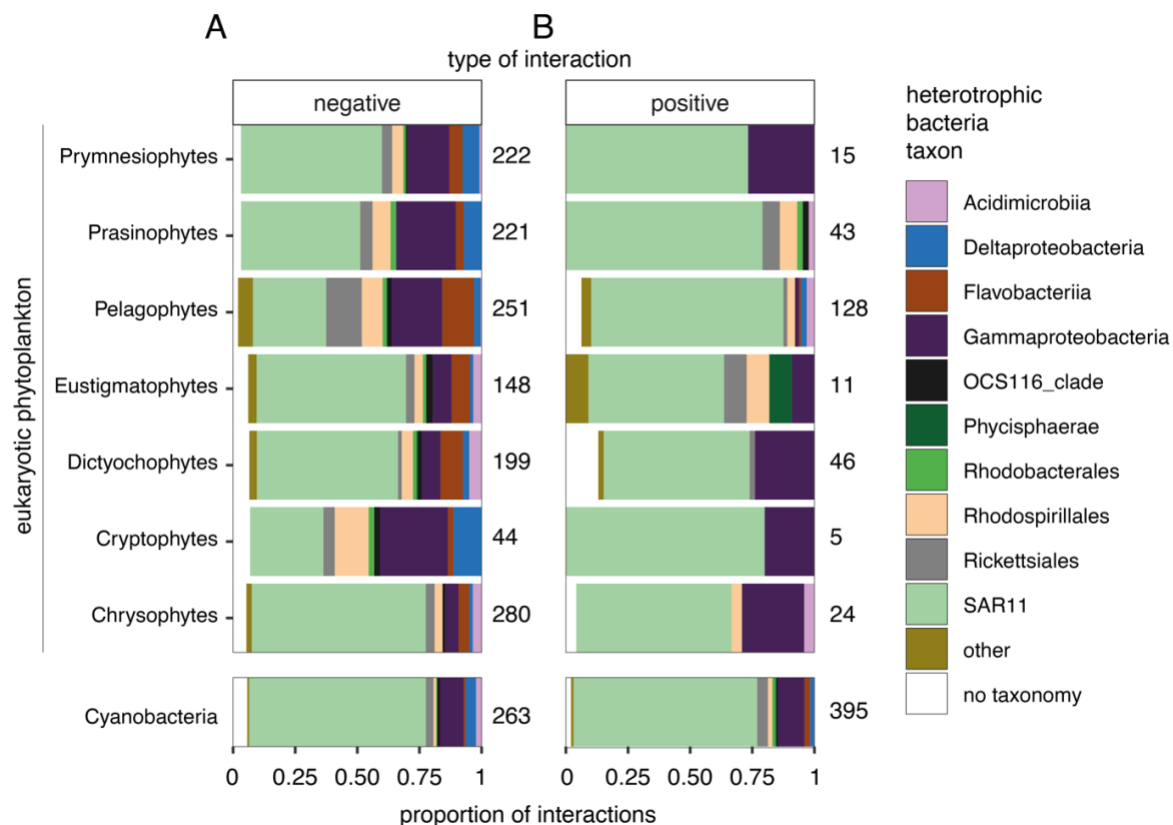
A taxonomic breakdown of negative and positive interactions revealed that across all bacterial taxa frequently detected within interactions, all lineages were generally involved in more negative than positive interactions (Figure 7E). SAR11, involved by far in the most positive and negative interactions is the most abundant and most diverse clade of heterotrophic bacteria observed at BATS, in line with its global dominance of the open ocean [72]. In combination with the requirement for organic material derived from phytoplankton, similar to other heterotrophs, this could explain the large number of interactions observed between SAR11 and phytoplankton ASVs. Furthermore, biogeography [73] and some work based on cultures of *Prochlorococcus* and SAR11 suggests that SAR11 relies on thiamin precursors produced by *Prochlorococcus* [74] and that there might be exchange of various molecules and co-evolution between them [75], which could explain the plethora of interactions observed between these two groups specifically. However, future exploration of interactions between *Prochlorococcus* and SAR11 is challenged by the complexity of

developing a co-culturing system [73]. A taxonomic breakdown of both positive and negative interactions by phytoplankton group enabled some insights into the various potential underlying reasons behind these interactions for different groups (Figure 7). Flavobacteriia (part of the Bacteroidota group) were almost exclusively involved in negative interactions across all phytoplankton lineages they were associated with. Bacteroidota are considered members of the phycosphere, living in close proximity with the phytoplankton cell, benefiting both the phytoplankton and bacterium in laboratory experiments [76]. Some members of the Bacteroidota are known algicidal bacteria (able to kill phytoplankton, and subsequently often found to benefit from and terminate algal blooms [77, 78]) suggesting that, even though the phycosphere is generally considered a microenvironment beneficial for the phytoplankton, some bacteria residing in the phycosphere could be preying on the phytoplankton cell, leading to a negative impact on phytoplankton abundance. Similar to Bacteroidota, Deltaproteobacteria were primarily involved in negative interactions across all taxa. Deltaproteobacteria have only been observed in association with the phycosphere in rare occasions and with generally low abundances, and no lineages have been isolated or investigated in depth [79]. Negative effects on phytoplankton have been reported before, but without any insights into the potential mechanism [80]. Planctomycetes from the Phycisphaerae lineage) show a different pattern, only being associated with eustigmatophytes in a positive interaction. Planctomycetes are often prevalent in algal blooms and are generally found in association with phytoplankton and algae surfaces [81, 82]. The Phycisphaerae specifically include multiple cultured representatives, isolated from the surface of marine macroalgae [83, 84]. Their positive association observed here suggest a beneficial impact on the phytoplankton, suggesting a specific association of the Phycisphaerae at BATS with the phycosphere of eustigmatophytes.

Gammaproteobacteria and Rickettsiales, as well as Rhodospirillales were found to frequently interact with phytoplankton, both negatively and positively, reflecting potentially diverse lifestyles within these taxa. Rickettsiales are known to vary in their host interaction greatly depending on host-type and their lineage within Rickettsiales [85]. While symbionts of the Rickettsiales lineage can live in mutualistic symbiosis with some hosts, very closely related lineages can have a pathogenic lifestyle within their host. This is further complexified by larger host ranges, with some species residing in a mutualistically beneficial symbiosis with their preferred host, while simultaneously switching to a pathogenic lifestyle when infecting another (see chapter 3). Less is known about Rhodospirillales, but they are known to be associated with phytoplankton and changes in phytoplankton DOM composition [86],

meaning that they are likely associated with the phycosphere. Our findings suggest that some members might have algicidal activities, potentially dependent on what phytoplankton they associate with.

Both the OCS116 clade and Acidimicrobiia seemed to have either positive or negative effects on phytoplankton based on the phytoplankton lineage. OCS116 is a clade of uncultured Alphaproteobacteria with an unknown lifestyle [87]. They are generally found in oligotrophic environments and tend to bloom during spring but are present throughout the whole year [88, 89]. Their rise in abundance during the spring indicates a response to the annual bloom in phytoplankton observed at BATS during the spring transition. Their positive association with prasinophytes, which are highest in abundance in spring as key elements of the short phytoplankton bloom observed annually [11], confirms their potential co-association. Acidimicrobiia had positive interactions with both prasinophytes and pelagophytes, while negatively impacting eustigmatophytes, dictyochophytes and Cyanobacteria.



**Figure 8: Relative taxonomic composition of the interactions between phytoplankton and heterotrophic bacteria.** Separate breakdowns for both (A) negative and (B) positive interactions. Note that this does not consider the relative abundance of the taxonomic lineages but the number of interactions between phytoplankton and heterotrophic bacteria only. The number of interactions per phytoplankton lineage and each combination are indicated on the right side of the panel.

Changes in the abundance of highly interactive eukaryotic phytoplankton over time causes changes in heterotrophic community

Notably, eight distinct chrysophytes demonstrated various interactions with heterotrophic bacteria, especially v1v2seq644 which interacted with 130 bacterial ASVs alone (Figure 7A, 7B), including 116 negative interactions. All chrysophytes with significant interactions showed increasing relative abundances at BATS over time. To investigate a possible effect that this rise in chrysophytes might have on the heterotrophic community in the surface at BATS, we investigated possible dynamics in the abundance of the bacterial ASVs interacting positively or negatively with chrysophytes. The network analysis and chrysophyte abundance changes over would suggest that bacteria negatively interacting with chrysophytes would decrease in relative abundance while those positively interacting would increase over time.

Indeed, over time, heterotrophic bacteria positively associated with chrysophytes increased in the stratified surface, just as their interacting phytoplankton did (Figure S6A). We did however not observe the opposite trend for ASVs negatively interacting with chrysophytes, with no significant changes observed over time (Figure S6B). We observed the same pattern for the most abundant dictyochophyte (v1v2seq261) and prymnesiophytes (v1v2seq223), both of which significantly increase in the stratified surface over time. v1v2seq261 had no positive interactions in our network, but the six ASVs it negatively interacted with showed no temporal trend in the summer season. v1v2seq223 had one positive interaction (with v1v2seq987), which showed a significant increase in the summer over time, while the 43 ASVs it interacted negatively with showed no temporal trend. Both dictyochophytes and prymnesiophytes also include strains that decrease over time significantly, but they did not occur in enough samples throughout the timeseries to be included in the network, and we therefore were unable to test if we would observe the opposite trends for decreasing taxa.

The positive trends we observed for eukaryotic phytoplankton lineages and the heterotrophic bacteria they positively interact with suggests that there is a beneficial relationship between the interacting taxa resulting in an increase of both over time. For example, some of the bacteria increasing over time could be beneficial symbionts associated with the phycosphere, and as the phytoplankton increases in relative abundance at the surface during the stratified season, those associated bacteria benefit and subsequently increase in relative abundance as well. There are likely other interactions occurring as well, as we see an increase in the relative abundance of positively associated SAR11 strains, a lineage of small free-living bacteria unlikely to attach to phytoplankton due to its well-established free-living

lifestyle [72]. Higher specificity in positive interactions could explain why we do not observe a negative trend over time for heterotrophic bacteria with negative interactions. Heterotrophic bacteria with positive associations to phytoplankton are often only associated with one ASV ( $1.39 \pm 1.13$  interactions per heterotrophic bacteria ASV), while heterotrophic bacteria with negative associations are often negatively associated to multiple different phytoplankton ASVs ( $2.99 \pm 3.40$  interactions). This significant difference ( $p < 0.0001$ ) potentially blurs signals in long-term trends for negative associations, while positive interactions, restricted to often only single phytoplankton ASVs result in a clear long-term trend observed herein (Figure S6).

#### General limitations of network analysis

One key limitation of this type of co-occurrence analysis lays in the detection of interactions involving rare lineages. The network analysis conducted herein, and comparable approaches generally require a minimum detection limit, meaning that ASVs not present in at least 20% of samples analysed are removed before computation of the network. As demonstrated in chapters 2 and 3, while globally present, symbionts of eukaryotes in general are often very rare, and can remain undetected during sequencing in a lot of samples. It is therefore likely that the frequency cut-off applied led to the loss of a lot of potential symbionts. This highlights the undetectability of symbiosis involving rare bacterial lineages as a major limitation of co-occurrence networks. We were unable to detect any of the symbioses discussed in previous chapters herein, despite the detection of both host and symbiont lineages at BATS, highlighting this limitation particularly for rare, yet diverse and widespread, symbionts of protists.

Bacteroidota lineages shown and discussed in chapter 2 as physically co-associated with various lineages of eukaryotic phytoplankton were shown to be the most abundant symbiont at BATS among lineages recovered from single cell sorts. In line with this, we did detect frequent interactions between Bacteroidota and various phytoplankton lineages at BATS, while fewer interactions were observed for less abundant Verrucomicrobiota and Planctomycetota. We were unable to detect significant interactions for the planctomycete lineages *Mariniblastus* or *Rubripirellula*, although we have identified them in physical co-association with phytoplankton (see chapter 2), and only identified a member of the Phycisphaerae interacting with Eustigmatophyceae. This could be partially explained by the exclusion of large fractions of the latter due to the frequency of occurrence cut-off required.

## Conclusions

Our analysis of dynamics in the surface at BATS suggest significant changes in the composition of the microbial community since 1991. We do observe a significant decline in species richness both in the total community and in eukaryotic phytoplankton, coinciding with major changes in temperature, nutrient concentrations, and nutrient inventory. Focusing on the stratified summer season, we identified a decline in both cyanobacteria and in eukaryotic phytoplankton abundance, relative to heterotrophic bacteria. Coinciding with a decline in eukaryotic phytoplankton is a relative increase in Chrysophytes, as well as a shift in prymnesiophytes towards mixotrophic lineages. This likely represents a response to a decline in nutrient inventory which mixotrophs compensate through predation. Bacteria that are specific in their interactions with phytoplankton respond by following the same temporal trends as the phytoplankton lineages they interact with, highlighting downstream effects of changes in the primary producers. Acceleration in warming observed at BATS will likely intensify a reduction in species richness and changes in community composition already taking place currently. Further analysis is required to investigate the downstream effects these changes will have on the microbial loop and carbon export.

### **Author Contributions**

F.W. conducted amplicon data analysis, computed the co-occurrence network and wrote the manuscript. F.W. and A.Z.W. conceptualized the analysis. S.J.G. and C.A.C. conducted and received funding for the timeseries sampling until 2013. N.B. and R.J. lead the overall BATS timeseries effort and monthly sampling outside of BIOS-SCOPE process cruises. K.L.V. processed raw sequencing data. A.Z.W. edited the manuscript and provided feedback. C.J.C. curated the initial version of the chrysophyte reference database. R.J.P., J.G. and K.Y. extracted DNA samples. K.L.V., R.J.P. and S.L. coordinated sequencing efforts and F.W., K.L.V, M.L.M, S.L., and J.C. prepared sequencing libraries. E.H., K.L., S.L., R.J.P., and R.C. curated timeseries metadata and R.C. created the physical framework. E.B.K., B.T., H.G.C., C.A.C., A.Z.W., S.J.G., L.B. and their respective groups contributed to the overall timeseries effort.

## References

1. Lee H, Calvin K, Dasgupta D, Krinner G, Mukherji A, Thorne P, et al. IPCC, 2023: Climate Change 2023: Synthesis Report, Summary for Policymakers. Contribution of Working Groups I, II and III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change [Core Writing Team, H. Lee and J. Romero (eds.)]. IPCC, Geneva, Switzerland. 2023.
2. Bopp L, Resplandy L, Orr JC, Doney SC, Dunne JP, Gehlen M, et al. Multiple stressors of ocean ecosystems in the 21st century: projections with CMIP5 models. *Biogeosciences* 2013; **10**: 6225–6245.
3. Kwiatkowski L, Aumont O, Bopp L, Ciais P. The Impact of Variable Phytoplankton Stoichiometry on Projections of Primary Production, Food Quality, and Carbon Uptake in the Global Ocean. *Global Biogeochem Cycles* 2018; **32**: 516–528.
4. Lomas MW, Bates NR, Johnson RJ, Steinberg DK, Tanioka T. Adaptive carbon export response to warming in the Sargasso Sea. *Nat Commun* 2022; **13**: 1211.
5. Kwiatkowski L, Torres O, Bopp L, Aumont O, Chamberlain M, Christian JR, et al. Twenty-first century ocean warming, acidification, deoxygenation, and upper-ocean nutrient and primary production decline from CMIP6 model projections. *Biogeosciences* 2020; **17**: 3439–3470.
6. Martinez E, Antoine D, D’Ortenzio F, Gentili B. Climate-Driven Basin-Scale Decadal Oscillations of Oceanic Phytoplankton. *Science (1979)* 2009; **326**: 1253–1256.
7. Boyce DG, Lewis MR, Worm B. Global phytoplankton decline over the past century. *Nature* 2010; **466**: 591–596.
8. Skyllas N, Bintanja R, Buma AGJ, Brussaard CPD, Gröger M, Hieronymus J, et al. Validation of stratification-driven phytoplankton biomass and nutrient concentrations in the Northeast Atlantic Ocean as simulated by EC-Earth. *Geosciences (Basel)* 2019; **9**: 450.
9. Doney SC, Ruckelshaus M, Emmett Duffy J, Barry JP, Chan F, English CA, et al. Climate Change Impacts on Marine Ecosystems. *Ann Rev Mar Sci* 2011; **4**: 11–37.
10. Giovannoni SJ, Vergin KL. Seasonality in Ocean Microbial Communities. *Science (1979)* 2012; **335**: 671–676.
11. Eckmann CA, Bachy C, Wittmers F, Strauss J, Blanco-Bercial L, Vergin KL, et al. Recurring seasonality exposes dominant species and niche partitioning strategies of open ocean picoeukaryotic algae. *Commun Earth Environ* 2024; **5**: 266.

12. Karl DM, Lukas R. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep Sea Research Part II: Topical Studies in Oceanography* 1996; **43**: 129–156.
13. Wickham H, Averick M, Bryan J, Chang W, McGowan LD, François R, et al. Welcome to the Tidyverse. *J Open Source Softw* 2019; **4**: 1686.
14. Sudek S, Everroad RC, Gehman A-LM, Smith JM, Poirier CL, Chavez FP, et al. Cyanobacterial distributions along a physico-chemical gradient in the Northeastern Pacific Ocean. *Environ Microbiol* 2015; **17**: 3692–3707.
15. Callahan BJ, McMurdie PJ, Rosen MJ, Han AW, Johnson AJA, Holmes SP. DADA2: High-resolution sample inference from Illumina amplicon data. *Nat Methods* 2016; **13**: 581–583.
16. Vergin KL, Beszteri B, Monier A, Cameron Thrash J, Temperton B, Treusch AH, et al. High-resolution SAR11 ecotype dynamics at the Bermuda Atlantic Time-series Study site by phylogenetic placement of pyrosequences. *ISME J* 2013; **7**: 1322–1332.
17. Choi CJ, Jimenez V, Needham DM, Poirier C, Bachy C, Alexander H, et al. Seasonal and Geographical Transitions in Eukaryotic Phytoplankton Community Structure in the Atlantic and Pacific Oceans. *Front Microbiol* 2020; **11**: 2187.
18. Strauss J, Choi CJ, Grone J, Wittmers F, Jimenez V, Makareviciute-Fichtner K, et al. The Bay of Bengal exposes abundant photosynthetic picoplankton and newfound diversity along salinity-driven gradients. *Environ Microbiol* 2023; **n/a**.
19. Katoh K, Standley DM. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* 2013; **30**: 772–780.
20. Capella-Gutierrez S, Silla-Martinez JM, Gabaldon T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* 2009; **25**: 1972–1973.
21. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* 2014; **30**: 1312–1313.
22. Czech L, Barbera P, Stamatakis A. Genesis and Gappa: Processing, analyzing and visualizing phylogenetic (placement) data. *Bioinformatics* 2020; **36**: 3263–3265.
23. Wang L-G, Lam TT-Y, Xu S, Dai Z, Zhou L, Feng T, et al. Treeio: An R Package for Phylogenetic Tree Input and Output with Richly Annotated and Associated Data. *Mol Biol Evol* 2020; **37**: 599–603.
24. Yu G, Smith DK, Zhu H, Guan Y, Lam TT-Y. ggtree: an r package for visualization and annotation of phylogenetic trees with their covariates and other associated data. *Methods Ecol Evol* 2017; **8**: 28–36.

25. Xu S, Dai Z, Guo P, Fu X, Liu S, Zhou L, et al. ggtreeExtra: Compact visualization of richly annotated phylogenetic data. *Mol Biol Evol* 2021.
26. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics* 2009; **10**: 1–9.
27. Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin P, O'Hara B, et al. Vegan: Community Ecology Package. *R Package Version 22-1* 2015; **2**: 1–2.
28. Hillebrand H, Blasius B, Borer ET, Chase JM, Downing JA, Eriksson BK, et al. Biodiversity change is uncoupled from species richness trends: Consequences for conservation and monitoring. *Journal of Applied Ecology* 2018; **55**: 169–184.
29. Wickham H. ggplot2: elegant graphics for data analysis. 2016. springer.
30. Anderson MJ. A new method for non-parametric multivariate analysis of variance. *Austral Ecol* 2001; **26**: 32–46.
31. Wood S, Wood MS. Package 'mgcv'. *R package version* 2015; **1**: 729.
32. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput Biol* 2012; **8**: e1002606-.
33. Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, et al. Determinants of community structure in the global plankton interactome. *Science (1979)* 2015; **348**: 1262073.
34. Deutschmann IM, Lima-Mendez G, Krabberød AK, Raes J, Vallina SM, Faust K, et al. Disentangling environmental effects in microbial association networks. *Microbiome* 2021; **9**: 232.
35. Schwager E, Weingart G, Bielski C, Huttenhower C. CCREPE: compositionality corrected by permutation and renormalization. *R/Bioconductor* <https://doi.org/10.18129/B.2014.9> 2014; **9**.
36. Csardi G, Nepusz T. The Igraph Software Package for Complex Network Research. *InterJournal* 2005; **Complex Systems**: 1695.
37. Csárdi G, Nepusz T, Traag V, Horvát S, Zanini F, Noom D, et al. igraph: Network Analysis and Visualization in R. 2023. URL <https://CRAN.R-project.org/package=igraph> R package version . , **1**
38. Weaver W. The mathematical theory of communication. 1963. University of Illinois Press.
39. Simpson EH. Measurement of Diversity. *Nature* 1949; **163**: 688.
40. Kuczynski L, Ontiveros VJ, Hillebrand H. Biodiversity time series are biased towards increasing species richness in changing environments. *Nat Ecol Evol* 2023; **7**: 994–1001.
41. Ustick LJ, Larkin AA, Garcia CA, Garcia NS, Brock ML, Lee JA, et al. Metagenomic analysis reveals global-scale patterns of ocean nutrient limitation. *Science (1979)* 2021; **372**: 287 LP – 291.

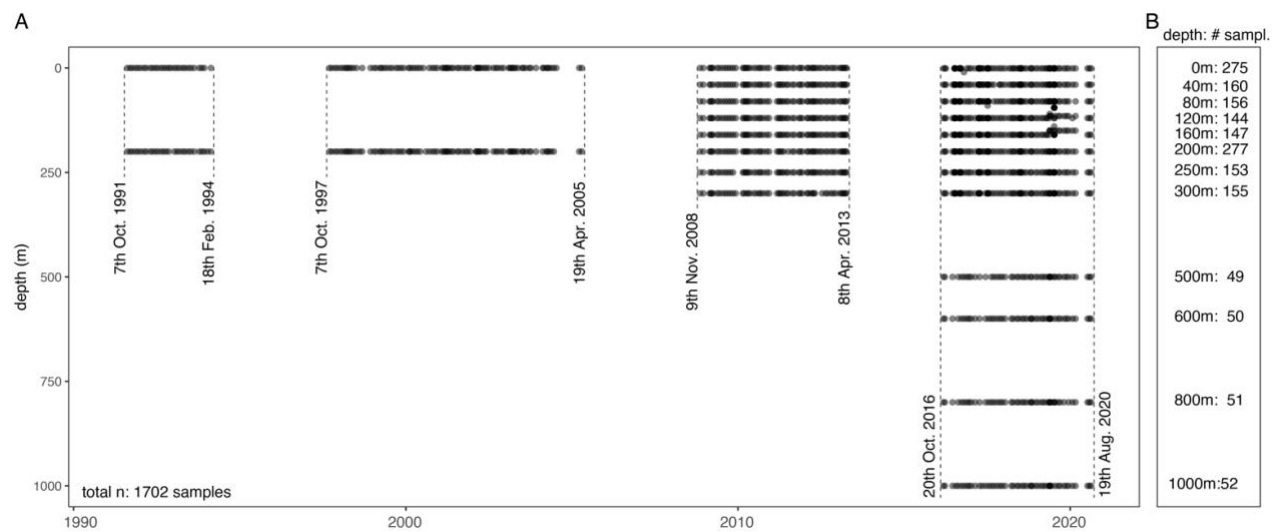
42. Malmstrom RR, Coe A, Kettler GC, Martiny AC, Frias-Lopez J, Zinser ER, et al. Temporal dynamics of *Prochlorococcus* ecotypes in the Atlantic and Pacific oceans. *ISME J* 2010; **4**: 1252–1264.
43. Chisholm SW, Olson RJ, Zettler ER, Goericke R, Waterbury JB, Welschmeyer NA. A novel free-living prochlorophyte abundant in the oceanic euphotic zone. *Nature* 1988; **334**: 340–343.
44. Grone J, Poirier C, Abbott K, Wittmers F, Jaeger GS, Mahadevan A, et al. A single *Prochlorococcus* ecotype dominates the tropical Bay of Bengal with ultradian growth. *Environ Microbiol* 2024; **26**: e16605.
45. Wei Y, Xuejin F, Ta-Hui L, Xingyu H, Le X, Shuzhen W, et al. Diverse Subclade Differentiation Attributed to the Ubiquity of *Prochlorococcus* High-Light-Adapted Clade II. *mBio* 2022; **13**: e03027-21.
46. Berube PM, Coe A, Roggensack SE, Chisholm SW. Temporal dynamics of *Prochlorococcus* cells with the potential for nitrate assimilation in the subtropical Atlantic and Pacific oceans. *Limnol Oceanogr* 2016; **61**: 482–495.
47. Moore LR, Rocap G, Chisholm SW. Physiology and molecular phylogeny of coexisting *Prochlorococcus* ecotypes. *Nature* 1998; **393**: 464–467.
48. Vaillancourt RD, Marra J, Seki MP, Parsons ML, Bidigare RR. Impact of a cyclonic eddy on phytoplankton community structure and photosynthetic competency in the subtropical North Pacific Ocean. *Deep Sea Research Part I: Oceanographic Research Papers* 2003; **50**: 829–847.
49. Siegel DA, McGillicuddy Jr DJ, Fields EA. Mesoscale eddies, satellite altimetry, and new production in the Sargasso Sea. *J Geophys Res Oceans* 1999; **104**: 13359–13379.
50. Eckmann CA, Eberle JS, Wittmers F, Wilken S, Bergauer K, Poirier C, et al. Eukaryotic algal community composition in tropical environments from solar salterns to the open sea. *Front Mar Sci* 2023; **10**.
51. Harbeitner RC, Wittmers F, Yung CCM, Eckmann CA, Hehenberger E, Blum M, et al. Gradients of bacteria in the oceanic water column reveal finely-resolved vertical distributions. *PLoS One* 2024; **19**: e0298139-.
52. Freilich MA, Poirier C, Dever M, Alou-Font E, Allen J, Cabornero A, et al. 3D intrusions transport active surface microbial assemblages to the dark ocean. *Proceedings of the National Academy of Sciences* 2024; **121**: e2319937121.
53. Seymour JR, Amin SA, Raina J-B, Stocker R. Zooming in on the phycosphere: the ecological interface for phytoplankton–bacteria relationships. *Nat Microbiol* 2017; **2**: 17065.

54. del Campo J, Massana R. Emerging Diversity within Chrysophytes, Choanoflagellates and Bicosoecids Based on Molecular Surveys. *Protist* 2011; **162**: 435–448.
55. Wilken S, Yung CCM, Poirier C, Massana R, Jimenez V, Worden AZ. Choanoflagellates alongside diverse uncultured predatory protists consume the abundant open-ocean cyanobacterium *Prochlorococcus*. *Proceedings of the National Academy of Sciences* 2023; **120**: e2302388120.
56. Lepère C, Vaultot D, Scanlan DJ. Photosynthetic picoeukaryote community structure in the South East Pacific Ocean encompassing the most oligotrophic waters on Earth. *Environ Microbiol* 2009; **11**: 3105–3117.
57. Nicholls KH, Wujek DE. 12 - CHRYSOPHYCEAN ALGAE. In: WEHR JD, SHEATH RG (eds). *Freshwater Algae of North America*. 2003. Academic Press, Burlington, pp 471–509.
58. Li Q, Edwards KF, Schvarcz CR, Steward GF. Broad phylogenetic and functional diversity among mixotrophic consumers of *Prochlorococcus*. *ISME J* 2022; **16**: 1557–1569.
59. Hartmann M, Zubkov M V, Scanlan DJ, Lepère C. In situ interactions between photosynthetic picoeukaryotes and bacterioplankton in the Atlantic Ocean: evidence for mixotrophy. *Environ Microbiol Rep* 2013; **5**: 835–840.
60. Wilken S, Yung CCM, Hamilton M, Hoadley K, Nzongo J, Eckmann C, et al. The need to account for cell biology in characterizing predatory mixotrophs in aquatic environments. *Philosophical Transactions of the Royal Society B: Biological Sciences* 2019; **374**: 20190090.
61. Flynn KJ, Stoecker DK, Mitra A, Raven JA, Glibert PM, Hansen PJ, et al. Misuse of the phytoplankton–zooplankton dichotomy: the need to assign organisms as mixotrophs within plankton functional types. *J Plankton Res* 2013; **35**: 3–11.
62. Hartmann M, Grob C, Tarran GA, Martin AP, Burkill PH, Scanlan DJ, et al. Mixotrophic basis of Atlantic oligotrophic ecosystems. *Proc Natl Acad Sci U S A* 2012; **109**: 5756–5760.
63. Dorrell RG, Azuma T, Nomura M, Audren de Kerdrel G, Paoli L, Yang S, et al. Principles of plastid reductive evolution illuminated by nonphotosynthetic chrysophytes. *Proceedings of the National Academy of Sciences* 2019; **116**: 6914–6923.
64. Wilken S, Choi CJ, Worden AZ. Contrasting Mixotrophic Lifestyles Reveal Different Ecological Niches in Two Closely Related Marine Protists. *J Phycol* 2020; **56**: 52–67.
65. Mitra A, Flynn KJ, Burkholder JM, Berge T, Calbet A, Raven JA, et al. The role of mixotrophic protists in the biological carbon pump. *Biogeosciences* 2014; **11**: 995–1005.
66. Ward BA, Follows MJ. Marine mixotrophy increases trophic transfer efficiency, mean organism size, and vertical carbon flux. *Proceedings of the National Academy of Sciences* 2016; **113**: 2958 LP – 2963.

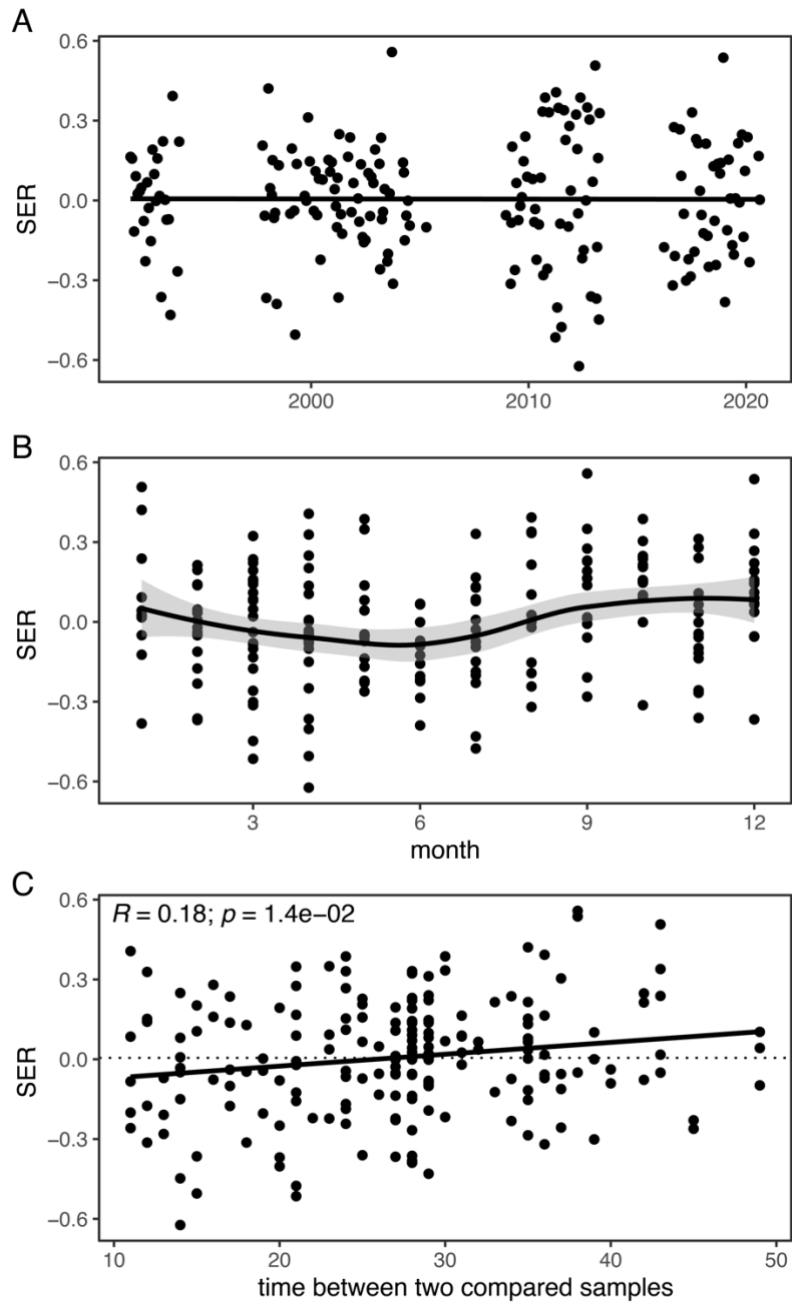
67. Blanco-Bercial L, Parsons R, Bolaños LM, Johnson R, Giovannoni SJ, Curry R. The protist community traces seasonality and mesoscale hydrographic features in the oligotrophic Sargasso Sea. *Front Mar Sci* 2022; **9**.
68. Cruz S, Cartaxana P. Kleptoplasty: Getting away with stolen chloroplasts. *PLoS Biol* 2022; **20**: e3001857-.
69. Zhu F, Massana R, Not F, Marie D, Vaultot D. Mapping of picoeucaryotes in marine ecosystems with quantitative PCR of the 18S rRNA gene. *FEMS Microbiol Ecol* 2005; **52**: 79–92.
70. Koppelle S, López-Escardó D, Brussaard CPD, Huisman J, Philippart CJM, Massana R, et al. Mixotrophy in the bloom-forming genus *Phaeocystis* and other haptophytes. *Harmful Algae* 2022; **117**: 102292.
71. Nygaard K, Tobiesen A. Bacterivory in algae: A survival strategy during nutrient limitation. *Limnol Oceanogr* 1993; **38**: 273–279.
72. Giovannoni SJ. SAR11 Bacteria: The Most Abundant Plankton in the Oceans. *Ann Rev Mar Sci* 2017; **9**: 231–255.
73. Becker JW, Hogle SL, Rosendo K, Chisholm SW. Co-culture and biogeography of *Prochlorococcus* and SAR11. *ISME J* 2019; **13**: 1506–1519.
74. Carini P, Campbell EO, Morré J, Sanudo-Wilhelmy SA, Cameron Thrash J, Bennett SE, et al. Discovery of a SAR11 growth requirement for thiamin's pyrimidine precursor and its distribution in the Sargasso Sea. *ISME J* 2014; **8**: 1727–1738.
75. Braakman R, Follows MJ, Chisholm SW. Metabolic evolution and the self-organization of ecosystems. *Proceedings of the National Academy of Sciences* 2017; **114**: E3091–E3100.
76. Jolley ET, Jones AK. The interaction between *Navicula muralis* grunow and an associated species of *Flavobacterium*. *British Phycological Journal* 1977; **12**: 315–328.
77. Paul C, Pohnert G. Interactions of the algicidal bacterium *Kordia algicida* with diatoms: regulated protease excretion for specific algal lysis. *PLoS One* 2011; **6**: e21032.
78. Coyne KJ, Wang Y, Johnson G. Algicidal bacteria: a review of current knowledge and applications to control harmful algal blooms. *Front Microbiol* 2022; **13**: 871177.
79. Chen Q, Zhu B, Sun D, Liu W, Sun X, Duan S. The effect of protocatechuic acid on the phycosphere in harmful algal bloom species *Scrippsiella trochoidea*. *Aquatic Toxicology* 2020; **227**: 105591.
80. Jin Z, Guo-Fu C, Ke-Zhen Y, Hui J, Jun-Ting S, Zhong-Hua C. Phycosphere Microbial Succession Patterns and Assembly Mechanisms in a Marine Dinoflagellate Bloom. *Appl Environ Microbiol* 2019; **85**: e00349-19.

81. Morris RM, Longnecker K, Giovannoni SJ. *Pirellula* and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. *Environ Microbiol* 2006; **8**: 1361–1370.
82. Faria M, Bordin N, Kizina J, Harder J, Devos D, Lage OM. Planctomycetes attached to algal surfaces: Insight into their genomes. *Genomics* 2018; **110**: 231–238.
83. Yoon J, Jang J-H, Kasai H. *Algisphaera agarilytica* gen. nov., sp. nov., a novel representative of the class Phycisphaerae within the phylum Planctomycetes isolated from a marine alga. *Antonie Van Leeuwenhoek* 2014; **105**: 317–324.
84. Fukunaga Y, Kurahashi M, Sakiyama Y, Ohuchi M, Yokota A, Harayama S. *Phycisphaera mikurensis* gen. nov., sp. nov., isolated from a marine alga, and proposal of Phycisphaeraceae fam. nov., Phycisphaerales ord. nov. and Phycisphaerae classis nov. in the phylum Planctomycetes. *J Gen Appl Microbiol* 2009; **55**: 267–275.
85. Driscoll TP, I VV, L GM, S LS, A RS, Magda B-S, et al. Wholly Rickettsia! Reconstructed Metabolic Profile of the Quintessential Bacterial Parasite of Eukaryotic Cells. *mBio* 2017; **8**: e00859-17.
86. Jain A, Meena RM, Naik RK, Gomes J, Bandekar M, Bhat M, et al. Response of polar front phytoplankton and bacterial community to micronutrient amendments. *Deep Sea Research Part II: Topical Studies in Oceanography* 2015; **118**: 197–208.
87. Suzuki MT, Rappé MS, Haimberger ZW, Winfield H, Adair N, Ströbel J, et al. Bacterial diversity among small-subunit rRNA gene clones and cellular isolates from the same seawater sample. *Appl Environ Microbiol* 1997; **63**: 983–989.
88. Morris RM, Frazar CD, Carlson CA. Basin-scale patterns in the abundance of SAR11 subclades, marine Actinobacteria (OM1), members of the Roseobacter clade and OCS116 in the South Atlantic. *Environ Microbiol* 2012; **14**: 1133–1144.
89. 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.

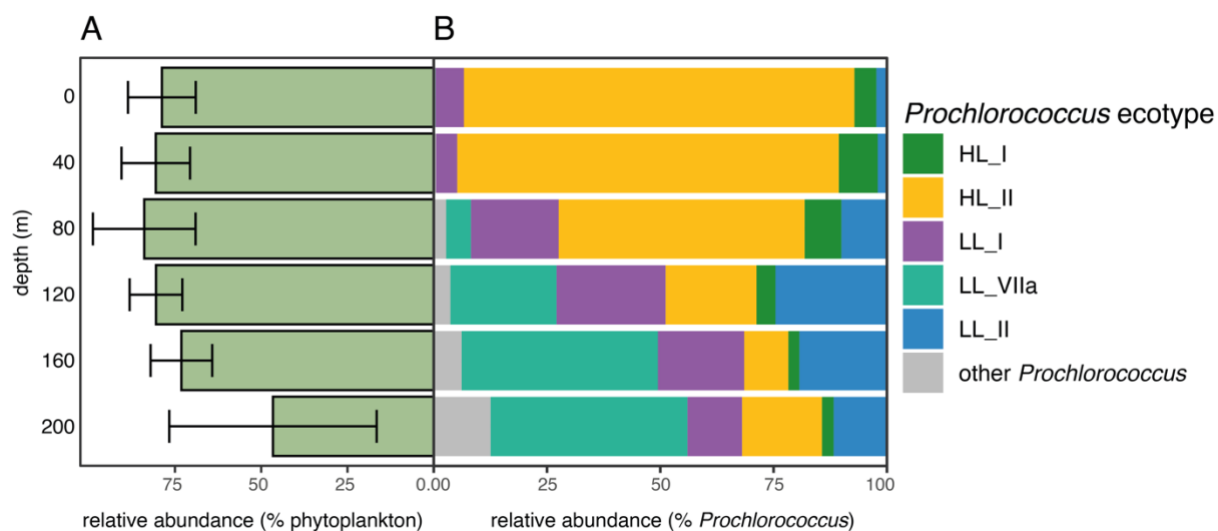
## Supplementary Figures



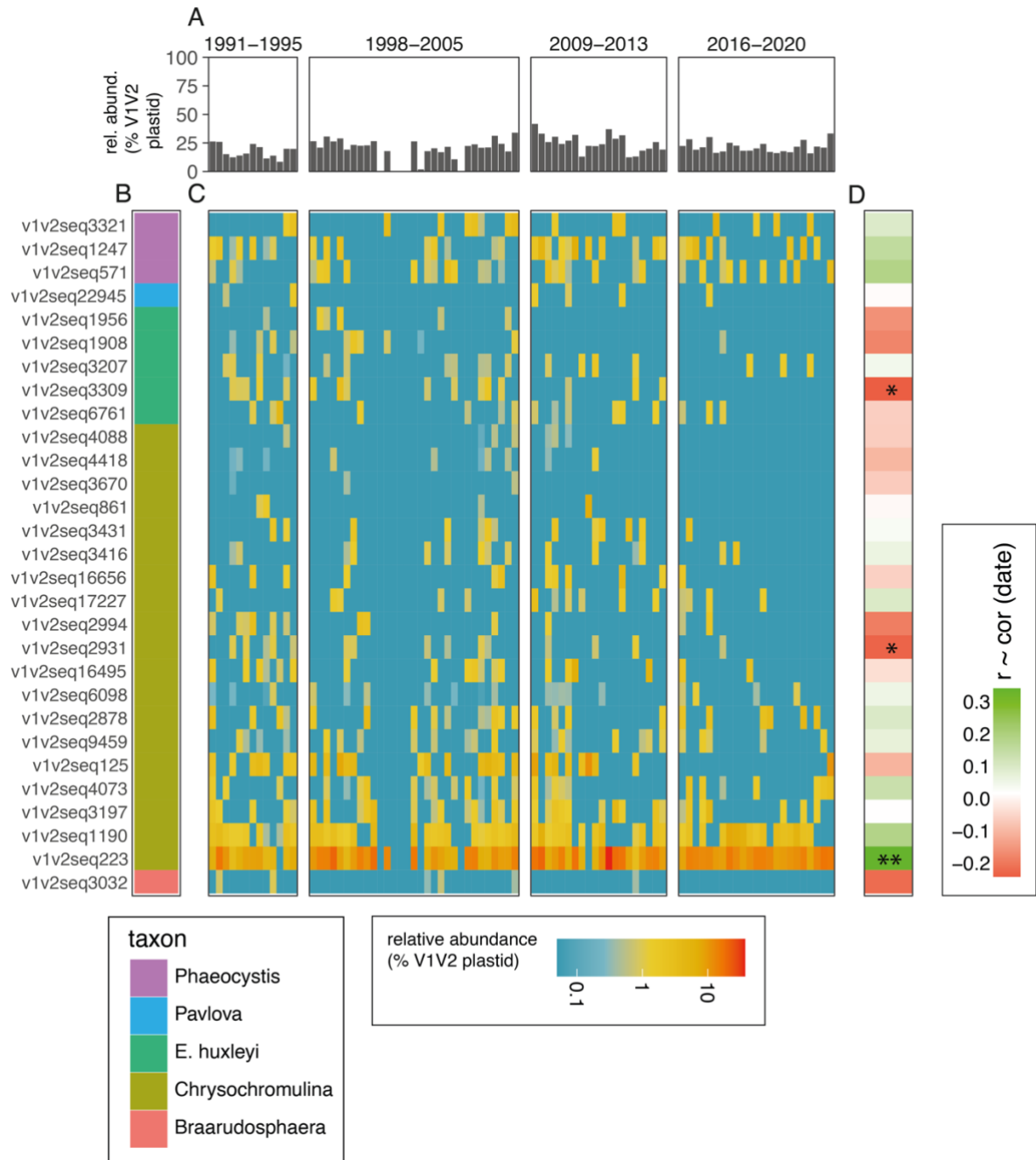
**Figure S1: Overview of molecular timeseries sampling at BATS between 1991 and 2020.** (A) sampling timepoints and depth-resolution of non-size-fractionated DNA collection. Sampling intensified with expansion of the general sampling-effort at BATS over time. The dataset comprises a total of 1702 amplicon samples. (B) Number of samples collected over the duration of the timeseries separated by depth. Initial sampling of the photic zone was initially restricted to the surface and expanded to multiple depth in 2008. Gaps in data are caused by the absence of any microbial timeseries sampling efforts.



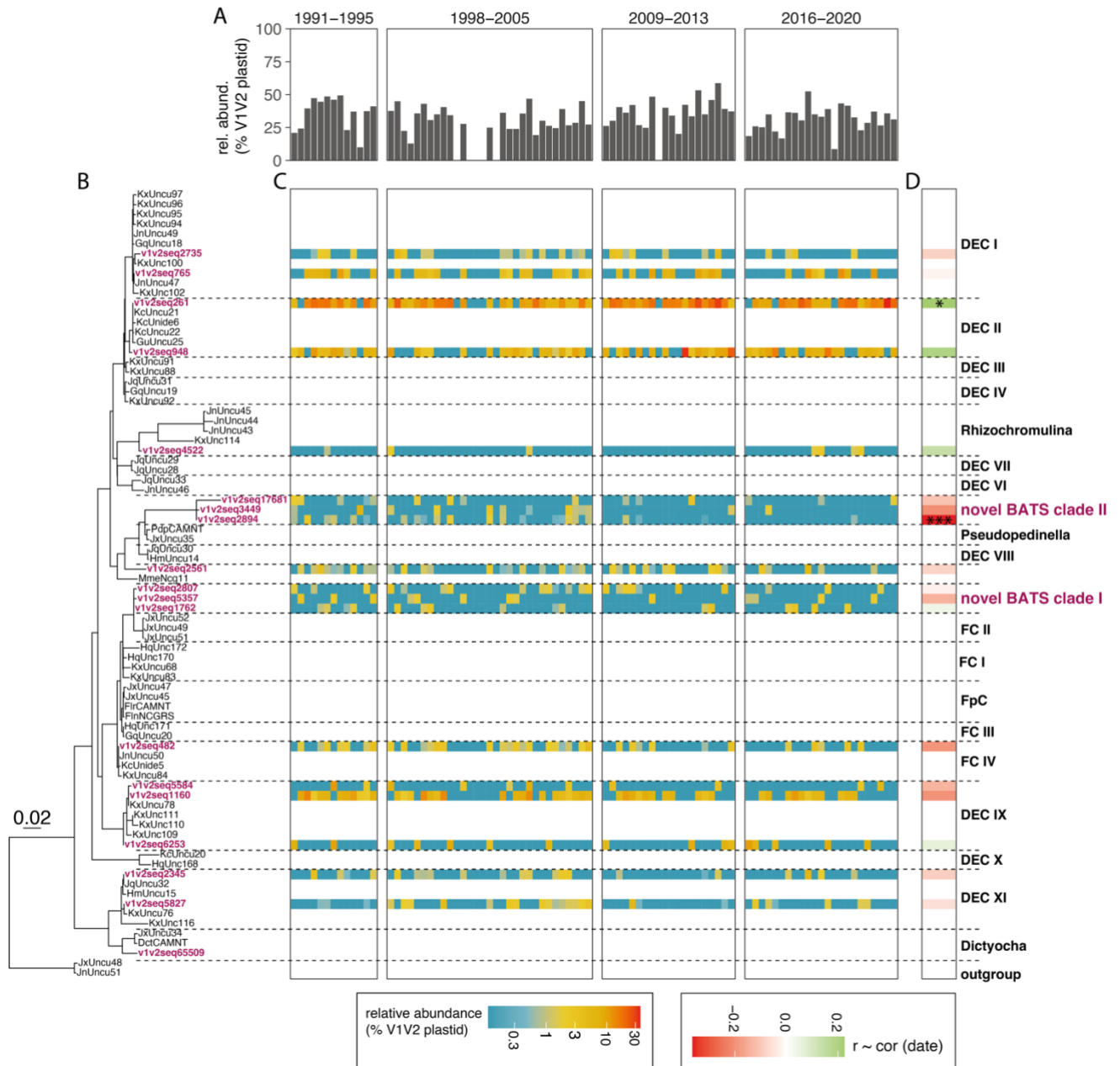
**Figure S2: Species exchange rate displays high variability.** (A) Change in SER over time throughout the BATS timeseries. (B) Relationship between SER and time between two sampling points. (C) SER depending on the time of sampling of the second sample used to compute the SER.



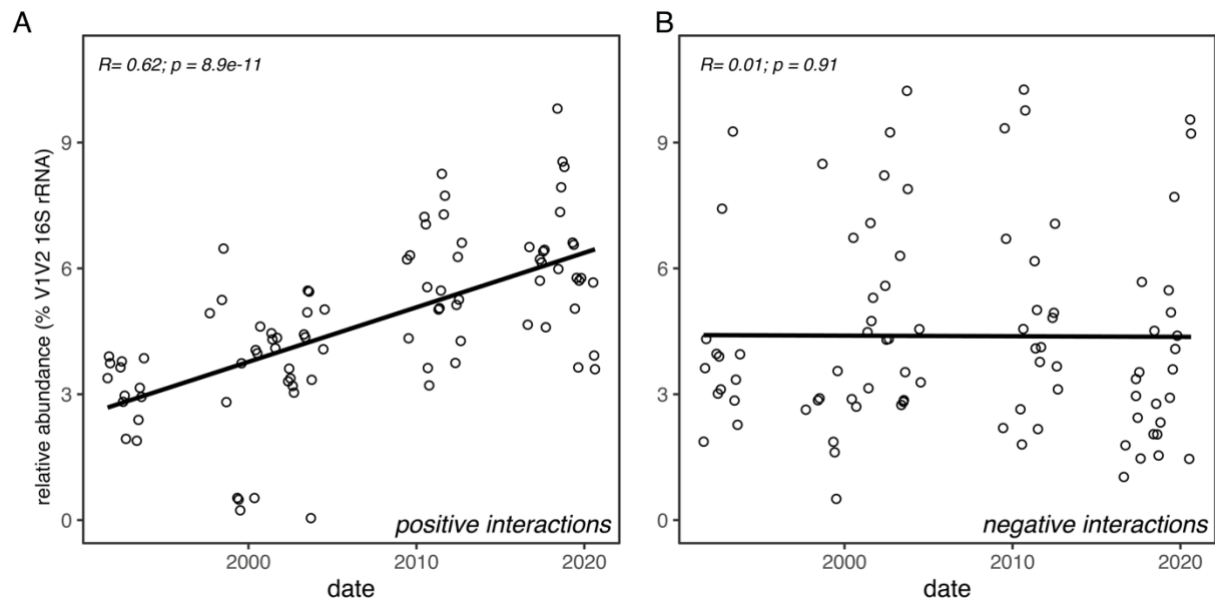
**Figure S3: Changes in *Prochlorococcus* abundance and ecotype composition over depth during the stratified season.** (A) Relative abundance of *Prochlorococcus* (all ecotypes) out of total phytoplankton (cyanobacteria and eukaryotic phytoplankton), averaged per depth. Error bars indicate standard deviations. (B) Relative contribution of the five most abundant ecotypes out of total *Prochlorococcus* across depth. All phytoplankton taxonomy and *Prochlorococcus* ecotype classifications are based on PhyloAssigner [11].



**Figure S4: Prymnesiophytes dynamics in the stratified season.** (A) Relative abundance of all prymnesiophytes out of V1V2 16S rRNA plastid reads (eukaryotic phytoplankton). (B) Prymnesiophyte taxonomy based on best homology to related sequences in NCBI. (C) Heatmap showing the relative abundance (out of eukaryotic phytoplankton) for each separate ASV per sample. (D) Correlation coefficient for correlation between the relative plastid abundance of each ASV with the date of sampling. Significant correlations are marked with via asterisks (\*  $p \leq 0.05$ ; \*\*  $p \leq 0.005$ ; \*\*\*  $p \leq 0.0005$ ). Years that included designated processing cruises during the summer stratified period were subsampled to reduce the number of samples. Only one randomly selected sample per cruise was considered.



**Figure S5: Environmental dictyochophyte clades respond to changing conditions differently.** (A) Relative abundance of all dictyochophytes out of V1V2 16S rRNA plastid reads (eukaryotic phytoplankton). (B) Maximum-likelihood phylogenetic reconstruction of dictyochophytes based on 16S rRNA gene sequences as published in [17]. The phylogeny was reconstructed under the GTR model with 1516 positions covering the V1V2 region. ASVs recovered herein and grafted onto the phylogeny based on best node phylogenetic placement are highlighted in purple. (C) Heatmap showing the relative abundance (out of eukaryotic phytoplankton) for each separate ASV per sample. (D) Correlation coefficient for correlation between the relative abundance out of plastid 16S rRNA and the date. Significant correlations are marked with via asterisks (\*  $p \leq 0.05$ ; \*\*  $p \leq 0.005$ ; \*\*\*  $p \leq 0.0005$ ). Years that included designated processing cruises during the summer stratified period were subsampled to reduce the number of samples. Only one sample per cruise was considered.



**Figure S6: Change in abundance of heterotrophic bacteria associated with chrysophytes.** Cumulative relative abundance (out of total V1V2 16S rRNA reads) of ASVs positively (A) or negatively (B) associated with chrysophyte ASVs in the BATS surface layer. Positive bacterial interactions with phytoplankton were more specific ( $1.39 \pm 1.13$  interactions per heterotrophic bacteria ASV) than negative interactions ( $2.99 \pm 3.40$  interactions).

## **Chapter 6**

### **Synthesis and Future Perspectives**

---

Fabian Wittmers

In summary, in this thesis I utilize two main methods to investigate microbial interactions, focused on associations between the biological domains of single-cell eukaryotes (protists) and heterotrophic bacteria. First, single cell-sorting and sequencing of photosynthetic and heterotrophic protists revealed clearly distinct microbial communities associated with eukaryotes of different lifestyles. We recovered a series of heterotrophic bacteria from the phycosphere of diverse eukaryotic phytoplankton, encoding a diverse repertoire of carbohydrate-active enzymes. Notably, this included pathways specialized in the breakdown of intact cell walls, including mechanisms to break protective bonds on the outside of phytoplankton cells, which suggests that some of these bacteria might not only utilize exudated carbon released into the phycosphere, but might also attack the eukaryotic cell wall itself. Heterotrophic protists had associated bacteria from three different phyla: the Margulisbacteria, previously not known to be associated with any protists, and the Rickettsiales and Coxiellales, both primarily known for their role as pathogens in animals, including humans. Through extensive phylogenetic reconstruction, we could establish the relationships of these marine relatives of human pathogens at phylogenetically important positions in the evolution of human pathogens. In the context of marine ecology, focusing on individually sorted cells revealed a variety of symbionts for uncultured lineages of heterotrophic protists, most of which had no known associated bacteria before (e.g., all MAST lineages discussed in chapters 3 and 4), suggesting that they might have a microbiome in some ways similar to that of their photosynthetic counterparts.

The single-cell component was expanded through the second approach, a long-term timeseries dataset from the Bermuda Atlantic Timeseries Study (BATS) site, utilized to create a co-occurrence network. Highlighting some of the long-term changes observed in the surface phytoplankton community at BATS over the last 30 years, the network analysis also revealed that a change in phytoplankton taxa does result in a corresponding response in their associated bacteria, when these bacteria are involved in lineage-specific interactions. In total, this work provides new insights into the interconnected microbes of the ocean, with implications in various niches that should be pursued further.

### **Broad importance of newly discovered marine protist-associated symbiont lineages**

Herein, we describe multiple novel lineages of symbionts associated with marine heterotrophic protists that are closely related with two lineages of widespread pathogens of animals, incl. livestock and humans, Coxiellales and Rickettsiales. This highlights the

economic and medical importance of these results. Our findings significantly expand the host range of the Coxiellales, previously only found in association with ticks and humans. The importance of reconstructing the evolution of these lineages and their host association has been acknowledged [1], but its marine origin highlighted herein (chapter 3) is often overlooked entirely. Part of the medical literature, by ignoring many of the basal marine lineages (partially because hosts are often unknown or no isolates exist), overlooks the transitions in lifestyle from free-living to host-associated to multiple independent transitions to multicellular hosts in both lineages. The fact that the Marinicoxiellaceae, closely related to the human pathogen *Coxiella burnetii*, associate with marine heterotrophic from the stramenopile MAST-3 lineage raises questions about which components of the genomes enable this vast diversity in host-range of close evolutionary relatives, highlighting the importance of further investigation. Although Coxiellales and Rickettsiales share many similarities in their lifestyle, they could be considered fundamentally different in their host-range, with Rickettsiales presenting a much more diverse range of host organisms spanning many eukaryotic supergroups, versus Coxiellales, only associated to animals as their primary host. Our findings indicate that this is to some degree and artefact of sampling-bias. We have a much better understanding of symbionts of ticks and other animals than their single-cell counterparts, and our findings, demonstrating a Coxiellales association with the stramenopile from the MAST-3 clade, and 16S rRNA surveys indicate that there is vast uncharacterized pool of symbionts related to these well-known animal symbionts in the marine environment. Exploration of this yet unknown marine diversity will be a key to the reconstruction of the evolution of these important symbionts and might help in our understanding on how these pathogens interact with their hosts.

### **Challenges in eukaryotic single-cell genomics**

The field of single-cell genomics, especially eukaryotic single-cell genomics, still faces a variety of significant challenges. First and foremost, generating eukaryotic genomes of sufficient quality for down-stream analysis of these genomes is currently very limited. Eukaryotic genomes are generally much larger than their prokaryotic counterparts, and often contain multiple copies of chromosomes, and often large regions with short sequence repeats or multiple copies of the same gene with an identical or near-identical sequence [2]. These factors complicate the assembly of eukaryotic genomes using currently applied assembly algorithms, with all of the commonly used ones relying on De Bruijn graphs as their

underlying fundamental concept [3, 4], which is known to struggle to decipher short multi-repeat regions or highly similar sequence fragments, resulting in low assembly qualities with short average sequence lengths. Although a large part of this thesis relies on eukaryotic single-cell sorts and subsequence sequencing of the sorted DNA, we were unable to utilize the eukaryotic DNA that was sequenced, but barely assembled into longer sequences, instead leading us to focusing on co-sorted prokaryotic organisms and their genomes. These genomes were generally much smaller, with fewer repeats and of better genome quality and completeness, enabling an in-depth annotation and analysis of their potential functionality, as well as genome-based reconstructions of their evolutionary position (see chapters 2-4). We were able to make predictions about potential pathways and other functions and had to rely on comparisons to related organisms with characterized isolates and assembled genomes. This limited absolute confirmation of specific interactions with their host, e.g., resistance to nutritional immunity through zinc sequestration by acquisition of transporters with zinc affinity (ZupT) in Protistiellales (see chapter 3). To be able to reconstruct not only the functions encoded by bacteria associated with protists, but their actual interactions on a genomic scale (complementing pathways for vitamin biosynthesis, etc.), will require advances in reconstructing the eukaryotic side of genome reconstruction from environmental data.

A considerable effort has been underway for decades, trying to improve the sequencing quality of environmental DNA, with the biggest breakthroughs and most promising advances being made through long-read sequencing [5, 6]. Traditional metagenomics sequencing relies on ‘shotgun’ sequencing, where DNA is split into small pieces, sequences, and then re-assembled. Long-read sequencing technologies can utilize higher quality DNA with longer, more intact genomic sections, sequencing these ‘long reads’ without prior splitting and are therefore much less susceptible to issues in reassembling regions containing extensive short repeats. These technologies are very promising for the field of eukaryotic genomics but are currently have two shortcomings. First, long-read sequencing requires much more DNA than short read sequencing [5, 6], a crucial limitation, especially in the context of single-cell genomics, where DNA concentrations are much lower than in conventional metagenomes. Second, long-read sequencing suffers from high error rates, which reduces sequencing data quality and induces accidental stop-codons through wrong base-calling. This can be compensated by conducting both short- and long-read sequencing on the same DNA, but this further increases the first problem mentioned above because this requires even more DNA. Currently, single-cell genomics and long-read sequencing are methods that are not

compatible, but long-read sequencing is continuously improving at rapid speed, requiring less and less DNA and with decreasing cost, are promising for potential future combination of these techniques.

### **Co-cultivation of protists is a necessary complement to culture-independent approaches**

This thesis relied entirely on culture-independent approaches, enabling the discovery of various symbionts of eukaryotic lineages that have no cultured representatives, allowing the description of the first symbionts of the MAST-3, MAST-4 and MAST-1 stramenopile lineages and an uncultured choanofagellate (*Bicosta minor*). This, however, also revealed some of the limits of culture-independent approaches such as single-cell sorting and sequencing, partially due to the inability to observe the actual interaction and relying almost completely on genomic sequences. Deciphering of mechanisms and their genomic basis in microbiology relies heavily on the ability to manipulate organisms, often by mutating genes, which is impossible with approaches applied herein. Especially considering the problems currently faced in the assembly of eukaryotic genomes, this has limitations in investigating symbiont-host relationships of these lineages in more depth. A lot of conclusions drawn about the functionality of the symbionts characterized genomically herein are drawn based on homology to symbionts that have been studied in interaction with their hosts in culture. This highlights the necessity of efforts in culturing more representatives of currently large groups of uncultured eukaryotic organisms, ideally together with their natural symbionts. Cultures are often maintained in axenic conditions, meaning, without bacteria, likely causing many symbionts to go unnoticed or being selected against during the isolation-process when using antibiotics to kill bacteria. Recently, groups have been able to culture increasingly more enigmatic groups of bacteria [7] and bacterial viruses [8], suggesting that more and more refined culturing techniques can contribute to culturing more representatives of uncultured or underrepresented protists, potentially with their natural microbiome intact. This would allow valuable insights into the nature of more of these interactions observed on a genomic scale herein.

### **Scales – implications of single-cell interactions for global pathways and the marketplace of chemical currencies**

Conceptually, in chemical oceanography, microbial activity creates and consumes a plethora of molecules, which drives an exchange of these products similar to a marketplace of

metabolites impacting global biogeochemical cycles [9, 10]. The field currently faces the challenge of measuring this exchange, that is, the production and release (through cell lysis, exudation, leakage, etc.) and uptake of these by all biological organisms, primarily microbes because it happens on extremely small timescales making these rates challenging to measure. Our findings herein highlight the spatial component that needs to be considered in these concepts. Microbes ‘see a sea of gradients’ [11], with highly localized environments surrounding their cells and a variety of concentration gradients existing on very small scales. For example, the phycosphere of phytoplankton (chapter 2) attracts a large variety of organisms, all of which utilize this microenvironment enriched in various dissolved carbon molecules. Some of these provide benefits for the phytoplankton in return, while others seem to have algicidal pathways encoded in their genomes and might attack the living phytoplankton cell. This creates a highly localized microenvironment booming with microbial activity and results in the rapid exchange of molecules between cells. The activity in this microenvironment is very different from the surrounding background not enriched by the phytoplankton, an example for metabolic heterogeneity on small scales. Similarly, heterotrophic protists seem to often be associated with bacteria (see chapters 3 & 4), both extracellularly and intracellularly, essentially harbouring their own microbiome, similar to phytoplankton organisms. This similarity to the phycosphere, although including fewer entities, creates a microenvironment in which symbiont and host likely exchange a plethora of molecules. As we understand these microenvironments better, the impact they have on the rates of production and uptake and subsequently the cycling of molecules on a larger scale crystallises as a question of ever increasingly complexity to investigate in the future.

### **Cell - to - cell interactions in co-occurrence network models**

Network analysis of phytoplankton and heterotrophic bacteria (chapter 5) showed extensive interactions, both positive and negative between the two groups. A comparison of these findings with phytoplankton-associated bacteria shows an overlap in some groups, especially those that are well established at BATS at relatively high abundances (e.g., Bacteroidota). Although not investigated, similar interactions likely exist within phytoplankton itself, especially considering the importance of putative mixotrophic protists observed in the surface ocean at BATS. Co-occurrence networks, as opposed to the cell-sorting approaches used herein also detect more variable interactions, e.g., predatory-prey interactions or those not involving physical attachment, while cell-sorting seems to only reveal actual physical

associations between co-sorted cells. But networks utilized herein (chapter 5), and their requirements reveal some of the major limitations that highlight the need for alternative approaches in the analysis of microbial interactions. The biggest shortcoming of co-occurrence networks lies in their occurrence cut-offs. For network models to be able to accurately assess the interaction between two entities, those need to be represented in a minimum number of samples investigated, typically ranging between 15 and 20%, depending on the exact model applied and, to some degree, personal preferences [12–14]. The work conducted herein using single-cell genomics revealed that most symbionts, although globally distributed, are very rare, almost never being detected at the frequency required for potential consideration in networks, limited by the sequencing depth of the amplicon samples. The new symbiotic lineages, as well as other known symbionts, are often rare when investigating a single variant, but are evolutionarily extremely diverse, and likely cumulatively abundant and important factors in shaping the microbial community and its impact on global biogeochemical cycles. Regardless of that, they are not being considered in current networks due to minimum occurrence cut-offs. Similar issues arise for interactions and species that, in the scenario of a timeseries underlying the interaction model, disappeared early on or appeared late within the timeseries, as well as highly seasonal and fluctuating components of the community. Accumulating for relative abundances of symbionts on higher taxonomic levels to be able to include them in network models is not an option due to their extreme variability in putative functionality and specialization on different host organisms. This highlights the necessity to employ different approaches from co-occurrence networks to explore microbial interactions, including, but not restricted to, single-cell sorting and co-cultivation efforts.

### **Long-term timeseries**

Timeseries studies have become increasingly popular with the decrease in sequencing costs for large datasets. But most molecular timeseries are collected in coastal settings, often off piers at research institutions (e.g., San Pedro Ocean Time-series location SPOT [15, 16]), due to challenging logistics of conducting high-resolution sampling in open ocean settings. Hawaii Ocean Timeseries Study (HOT) and Bermuda Atlantic Timeseries Study (BATS) in the subtropical Pacific and western North Atlantic open ocean being are the two major exceptions to this, officially established in 1988 [17, 18]. Long-term molecular timeseries studies provide unique insights and observations into the effects of climate change and

anthropogenic actions, especially when running for long time periods. The dataset analysed herein is unique in that its molecular record starts in 1991, earlier than any other open ocean timeseries, providing an important baseline for exploration of future change. Maintaining these long-term sampling efforts is of essential importance to monitor future changes in the microbial community, especially considering the acceleration in warming observed at BATS since 2010. However, it is important to note that, e.g., BATS, started in 1991, does not reflect a baseline of the microbial community not affected by anthropogenic climate change. The concept of ‘shifting baselines’, developed in fisheries science to describe how each generation of scientists accepts the state of the ecosystem at the beginning of their career as the natural baseline [19]. does apply to this situation as well. Maintaining the longest running timeseries, with as little interruption as possible, counteracts this ‘shifting baseline syndrome’ and should be prioritized in the future, to avoid a continuous drift in what we consider the ‘normal’ state of the microbial community, a state that, in reality, reflects a community altered by extensive anthropogenic forcing for decades already.

## References

1. Duron O, Doublet P, Vavre F, Bouchon D. The importance of revisiting Legionellales diversity. *Trends Parasitol* 2018; **34**: 1027–1037.
2. Harriet A, K HS, I KA, Maria P, J TB, J NC, et al. Eukaryotic genomes from a global metagenomic data set illuminate trophic modes and biogeography of ocean plankton. *mBio* 2023; **14**: e01676-23.
3. Bankevich A, Nurk S, Antipov D, Gurevich AA, Dvorkin M, Kulikov AS, et al. SPAdes: A New Genome Assembly Algorithm and Its Applications to Single-Cell Sequencing. *Journal of Computational Biology* 2012; **19**: 455–477.
4. Li D, Liu C-M, Luo R, Sadakane K, Lam T-W. MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics* 2015; **31**: 1674–1676.
5. Jain M, Olsen HE, Paten B, Akeson M. The Oxford Nanopore MinION: delivery of nanopore sequencing to the genomics community. *Genome Biol* 2016; **17**: 1–11.
6. Rhoads A, Au KF. PacBio Sequencing and its Applications. *Genomics Proteomics Bioinformatics* 2015; **13**: 278–289.
7. Lim Y, Seo J-H, Giovannoni SJ, Kang I, Cho J-C. Cultivation of marine bacteria of the SAR202 clade. *Nat Commun* 2023; **14**: 5098.
8. Buchholz HH, Michelsen ML, Bolaños LM, Browne E, Allen MJ, Temperton B. Efficient dilution-to-extinction isolation of novel virus–host model systems for fastidious heterotrophic bacteria. *ISME J* 2021; **15**: 1585–1598.
9. Moran MA, Kujawinski EB, Schroer WF, Amin SA, Bates NR, Bertrand EM, et al. Microbial metabolites in the marine carbon cycle. *Nat Microbiol* 2022; **7**: 508–523.
10. Wienhausen G, Noriega-Ortega BE, Niggemann J, Dittmar T, Simon M. The exometabolome of two model strains of the Roseobacter group: a marketplace of microbial metabolites. *Front Microbiol* 2017; **8**: 1985.
11. Stocker R. Marine Microbes See a Sea of Gradients. *Science (1979)* 2012; **338**: 628 LP – 633.
12. Deutschmann IM, Lima-Mendez G, Krabberød AK, Raes J, Vallina SM, Faust K, et al. Disentangling environmental effects in microbial association networks. *Microbiome* 2021; **9**: 232.
13. Lima-Mendez G, Faust K, Henry N, Decelle J, Colin S, Carcillo F, et al. Determinants of community structure in the global plankton interactome. *Science (1979)* 2015; **348**: 1262073.

14. Faust K, Sathirapongsasuti JF, Izard J, Segata N, Gevers D, Raes J, et al. Microbial Co-occurrence Relationships in the Human Microbiome. *PLoS Comput Biol* 2012; **8**: e1002606-.
15. Yeh Y-C, Fuhrman JA. Contrasting diversity patterns of prokaryotes and protists over time and depth at the San-Pedro Ocean Time series. *ISME Communications* 2022; **2**: 36.
16. Needham DM, Fuhrman JA. Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nat Microbiol* 2016; **1**: 1–7.
17. Karl DM, Lukas R. The Hawaii Ocean Time-series (HOT) program: Background, rationale and field implementation. *Deep Sea Research Part II: Topical Studies in Oceanography* 1996; **43**: 129–156.
18. Bates NR, Michaels AF, Knap AH. Seasonal and interannual variability of oceanic carbon dioxide species at the US JGOFS Bermuda Atlantic Time-series Study (BATS) site. *Deep Sea Research Part II: Topical Studies in Oceanography* 1996; **43**: 347–383.
19. Pauly D. Anecdotes and the shifting baseline syndrome of fisheries. *Trends Ecol Evol* 1995; **10**: 430.