Molecular ecological characterisation of high-latitude bacterioplankton

Dissertation
zur Erlangung des Grades eines
Doktors der Naturwissenschaften
- Dr. rer. nat. -

dem Fachbereich Biologie/Chemie
der Universität Bremen
vorgelegt von

Taylor Priest

Bremen
October 2022

Die vorliegende Doktorarbeit wurde im Rahmen des Programms *International Max Planck Research School of Marine Microbiology* (MarMic) in der Zeit von April 2019 bis October 2022 am Max Planck Institut für Marine Mikrobiologie angefertigt.

This thesis was prepared under the framework of the *International Max Planck Research School of Marine Microbiology* (MarMic) at the Max Planck Institute for Marine Microbiology from April 2019 to October 2022.

Gutachter: PD Dr. Bernhard M. Fuchs
Gutachter: Prof. Dr. A. Murat Eren
Gutachterin: Dr. Silva González Acinas

Prüfer: Prof. Dr. Jan-Hendrik Hehemann
Prüfer: PD Dr. Bernhard M. Fuchs
Prüfer: Prof. Dr. A. Murat Eren
Prüferin: Dr. Silva González Acinas

Datum des Promotionskolloquiums: 28.11.2022

Table of Contents

Summary

The Arctic Ocean is undergoing irreversible perturbations as a result of accelerated climate warming. Of major significance is the expanding influence of Atlantic water that expedites sea-ice decline, alters stratification and vertical mixing of the water column and facilitates northward expansion of temperate biota. Our understanding on how these processes will impact biological communities is severely limited. The Fram Strait is the primary entry route for Atlantic water into the Arctic Ocean and exit point for polar water and sea-ice. With the presence of two major current systems combined with horizontal mixing processes, the Fram Strait is characterised by a longitudinal gradient of hydrographic regimes reflective of Arctic, mixed and Atlantic conditions. This provides an invaluable opportunity to study the ecology of microbes over an environmental gradient and under changing conditions. Furthermore, given its high-latitude position, it also facilitates investigations on how dramatic seasonal transformations in conditions, such as sea-ice cover and light availability, influence microbes in the context of water mass history. This thesis provides an ecological characterisation of microbial communities over temporal and spatial scales in the Fram Strait in an effort to address these topics.

In Chapter II, we employed metagenomics from short- and long-read sequencing platforms to gain insights into microbial community composition across water masses in the Fram Strait. As that study incorporated the first PacBio HiFi (long-read) metagenomes from the marine environment, it was necessary to perform a methodological comparison. We show that using PacBio HiFi metagenomes, we are able to recover more metagenome-assembled genomes (MAGs) that, on average, are more complete, less fragmented and more frequently contain complete rRNA gene operons compared to using short-read metagenomes. This not only influenced our investigative toolkit throughout the remainder of this thesis but provides valuable data for future considerations on using long-read metagenomics in the study of marine microbial ecology.

From the analysis conducted in Chapter II, we observed a flavobacterial clade that is commonly associated with coastal temperate ecosystems, the NS5 Marine Group, to be prominent in high-latitude waters. This motivated us to delve deeper into this group and understand their diversity and function. By combining cultivation, metagenomics, epifluorescence and transmission electron microscopy, we were able to delineate this group into four novel candidate genera and evidence distinctions in function and spatiotemporal dynamics at the species and genus level (Chapter III). In that study, we also presented the first pure isolate and complete genome for a member of the NS5 Marine Group.

In Chapter IV, we performed the first high-resolution temporal analysis on microbial taxonomy and function in Arctic polar waters. Using a four-year 16S amplicon dataset and one

annual cycle of PacBio HiFi metagenomes, we evidenced that Atlantic water influx and sea-ice cover had a profound impact on the composition and function of microbial communities. Based on their omnipresence irrespective of conditions, we also identified a small fraction of the community that likely represents the resident microbiome of the Fram Strait. Furthermore, we showed that a transition to low-ice and high Atlantic water influx shifted the community to one dominated by heterotrophic clades that are functionally linked to phytoplankton-derived organic matter. Our findings suggest that the continued expansion of Atlantic water into the Arctic Ocean will be reflected in a Biological Atlantification of the microbial community, with populations adapted to Arctic conditions exhibiting reduced ecological niche space. These changes will have implications for the future ecosystem functioning and the carbon cycle.

In Chapter V of this thesis, we combined metagenomics and metatranscriptomics with analytical techniques to characterise the carbohydrate fraction of particulate organic matter and carbohydrate utilisation by microbes in the Atlantic waters of the Fram Strait during late summer. A high spatial heterogeneity was observed in both carbohydrates and their utilisation, which indicated patchiness in local productivity and a responsive microbial community. Carbohydrate utilisation was dominated by distinct microbial assemblages across sampling sites and consisted of populations making use of labile (communal) and more complex (specialist) substrates. We therein proposed that local biological and physical processes are important for continuing to shape the availability and utilisation of carbohydrates into the late summer.

In an effort to clearly and concisely convey the main findings from this thesis in the context of its original aims, a detailed description on the current and future state of the Fram Strait and Arctic Ocean microbiome is provided in the discussion. In addition, insights and recommendations on how to apply long-read metagenomes to answer questions on microbial ecology is provided, given its fundamental importance for this thesis and its relative infancy in environmental research applications. Lastly, owing to it representing an underlying theme throughout much of the research conducted, a discussion on the ecological niche concept is provided along with a proposal for its redefinition in marine microbial ecology.

Zusammenfassung

Der Arktische Ozean ist infolge der beschleunigten Klimaerwärmung irreversiblen Veränderungen ausgesetzt. Von großer Bedeutung ist der zunehmende Einfluss des atlantischen Wassers, der den Rückgang des Meereises beschleunigt, die Zonierung und vertikale Durchmischung der Wassersäule verändert und die Ausbreitung Arten der gemäßigten Breiten nach Norden erleichtert. Bisher wissen wir wenig darüber, wie sich diese Prozesse auf die Lebensgemeinschaften auswirken werden. Die Framstraße ist die wichtigste Verbindung für atlantisches Wasser um in den Arktischen Ozean zu gelangen und der Austrittspunkt für polares Wasser und Meereis. Durch die zwei großen Strömungssystemen in Kombination mit horizontalen Vermischungsprozessen, ist die Framstraße durch einen Längsgradienten von hydrographischen Systemen gekennzeichnet, welch arktische, gemischte und atlantische Bedingungen widerspiegeln. Dies bietet eine unschätzbare Gelegenheit, die Ökologie von Mikroben über eben diese Umweltgradienten und unter wechselnden Bedingungen zu untersuchen. Darüber hinaus ermöglicht die Lage in den hohen Breitengraden die Untersuchung, wie dramatische jahreszeitliche Veränderungen, z. B. die Meereisbedeckung und Lichtverfügbarkeit, die Mikroben beeinflussen unter Berücksichtigung der Herkunft der Wassermassen. Unter diesen Gesichtspunkten wird in dieser Arbeit eine ökologische Charakterisierung der mikrobiellen Gemeinschaften in der Framstraße über zeitliche und räumliche Skalen hinweg vorgenommen.

In Kapitel II haben wir Metagenomik mit Hilfe von „Short“- und „Long-Read“-Sequenzierungsplattformen eingesetzt, um Einblicke in die Zusammensetzung der mikrobiellen Gemeinschaften in den verschiedenen Wassermassen der Framstraße zu gewinnen. Da in dieser Studie die ersten PacBio HiFi („long-read“) Metagenome von marinen Proben verwendet wurden, war es notwendig, einen methodischen Vergleich durchzuführen. Wir zeigen, dass mit PacBio HiFi-Metagenomen mehr Metagenom assemblierte Genome (MAGs) gewonnen werden können, die im Durchschnitt vollständiger und weniger fragmentiert sind und häufiger vollständige rRNA-Gen-Operone enthalten als bei der Verwendung von „Short-Read“-Metagenomen. Dies beeinflusste nicht nur unser Vorgehen im weiteren Verlauf dieser Arbeit, sondern liefert auch wertvolle Daten für künftige Überlegungen zur Verwendung von „Long-Read“-Metagenomen bei der Erforschung der mikrobiellen Ökologie des Meeres.

Bei der in Kapitel II durchgeführten Analyse stellten wir fest, dass eine Gruppe der Flavobakterien, die üblicherweise mit Ökosystemen der gemäßigten Breiten assoziiert wird, die „NS5 Marine Group“, in Gewässern der hohen Breiten weit verbreitet ist. Dies motivierte uns diese Gruppe genauer zu untersuchen und ihre Vielfalt und Funktion zu verstehen. Durch

die Kombination von Kultivierung, Metagenomik, Epifluoreszenz- und Transmissionselektronenmikroskopie konnten wir diese Gruppe in vier neue Kandidatengattungen unterteilen und Unterschiede in der Funktion, sowie in der Raum-Zeit Dynamik auf der Ebene der Arten und Gattungen nachweisen (Kapitel III). Im Rahmen dieser Studie haben wir auch das erste reine Isolat und vollständige Genom eines Mitglieds der „NS5 Marine Group“ vorgestellt.

In Kapitel IV führten wir die erste hochauflösende zeitliche Analyse der mikrobiellen Taxonomie und Funktion in arktischen Polargewässern durch. Anhand eines vierjährigen 16S-Amplikon-Datensatzes und eines jährlichen Zyklus von PacBio HiFi-Metagenomen konnten wir nachweisen, dass der Zustrom von Atlantikwasser und die Meereisbedeckung einen weitreichenden Einfluss auf die Zusammensetzung und Funktion der mikrobiellen Gemeinschaften haben. Da sie beständig zugegen, konnten wir auch einen kleinen Teil der Gemeinschaft identifizieren, der wahrscheinlich das Kern-Mikrobiom der Framstraße darstellt. Darüber hinaus konnten wir zeigen, dass der Übergang zu Eis armen Verhältnissen und hohem Wasserzufluss aus dem Atlantik zu einer Gemeinschaft führte, die von heterotrophen Gruppen dominiert wird. Diese Heterotrophen wiederum sind funktionell mit organischem Material aus dem Phytoplankton verbunden. Unsere Ergebnisse deuten darauf hin, dass sich die fortgesetzte Ausdehnung des Atlantikwassers in den Arktischen Ozean in einer biologischen Atlantifizierung der mikrobiellen Gemeinschaft widerspiegeln wird, wobei Populationen, die an die arktischen Bedingungen angepasst sind, weniger ökologische Nischenraum finden. Diese Veränderungen werden sich auf das zukünftige Funktionieren des Ökosystems und den Kohlenstoffkreislauf auswirken.

In Kapitel V dieser Arbeit haben wir Metagenomik und Metatranskriptomik mit analytischen Techniken kombiniert, um die Kohlenhydrate von partikulärem organischem Material und die Kohlenhydrate, welche durch Mikroben verwertet werden, in den atlantischen Gewässern der Framstraße im Spätsommer zu charakterisieren. Sowohl bei den Kohlenhydraten als auch bei ihrer Verwertung wurde eine große räumliche Heterogenität festgestellt, was auf eine uneinheitliche lokale Produktivität und eine reaktionsfähige mikrobielle Gemeinschaft hindeutet. Die Verwertung der Kohlenhydrate wurde an den verschiedenen Probenahmestellen von unterschiedlichen Zusammensetzungen an Mikroben dominiert und bestand aus Populationen, die labile (gemeinschaftliche) und komplexere (spezialisierte) Substrate nutzten. Wir schlagen darin vor, dass lokale biologische und physikalische Prozesse wichtig sind, um die Verfügbarkeit und Nutzung von Kohlenhydrate bis in den Spätsommer hinein zu beeinflussen.

Um die wichtigsten Ergebnisse dieser Arbeit im Kontext ihrer ursprünglichen Ziele klar und prägnant zu vermitteln, wird in der Diskussion eine detaillierte Beschreibung des aktuellen und zukünftigen Zustands des Mikrobioms der Framstraße und des Arktischen Ozeans

gegeben. Darüber hinaus werden Einsichten und Empfehlungen zur Anwendung von „Long-Read“-Metagenomen zur Beantwortung von Fragen zur mikrobiellen Ökologie gegeben, da diese für diese Arbeit von grundlegender Bedeutung sind und in der Umweltforschung erst in relativ geringem Umfang eingesetzt werden. Schließlich wird das Konzept der ökologischen Nische diskutiert und ein Vorschlag für seine Neudefinition in der marinen mikrobiellen Ökologie unterbreitet, da es ein grundlegendes Thema für einen Großteil der durchgeführten Forschung darstellt.

Introduction

Since the emergence of life 3.6 Ga, the oceans have been teeming with microbes. Microbes were responsible for transforming Earth's conditions to facilitate the evolution of more complex life forms and have continued to shape and maintain ocean ecosystems ever since. In the current day, it is estimated that there are more microbes in the oceans than stars in the universe and collectively, they constitute more than 70% of the living biomass [1]. Within each litre of seawater, there can be billions of microbes comprised of tens of thousands of species that harbour myriad functions and metabolisms. This diversity underpins their ecological and biogeochemical significance. Notwithstanding the essential functions they perform for the maintenance of ocean ecosystems, marine microbes also play fundamental roles in mediating global biogeochemical cycles and are major influencers of climate [2, 3]. In order to understand ecosystem health and functioning, it is thus essential that we understand microbial ecology.

The field of marine microbial ecology is tasked with characterising the diversity, abundance and function of microbes and understanding their interactions with the biotic and abiotic components of their environment. At its core, there exist three fundamental questions:

1) What microbes are there?
2) How many are there?
3) What are they doing?

Despite the simple nature of these questions, the small size of microbes combined with the immense scale of the marine environment present major challenges. Research in marine microbial ecology is thus reliant on specialised techniques. Concurrent with technological developments over the past 50 years, our understanding on the diversity of microbes in the oceans, their metabolisms, ecological roles and dynamics has improved tremendously. However, there remain many unanswered questions and gaps in our knowledge. In particular, with respect to high-latitude waters and polar oceans, which remain under-sampled compared to temperate and tropical ecosystems. Considering that they account for 20% of the world's ocean surface, are the engines of the thermohaline circulation [4] and regulators of global climate [5], developing a deeper understanding of polar ocean ecosystems is of great importance.

1.1 Arctic Ocean

The Arctic Ocean is the smallest but also the most unique of the world's oceans. Spanning an area of 6 million km^2, it comprises only 1% of global ocean volume yet contains 15% of the world's shelf seas [6]. It is the only ocean in which continental shelfs comprise >50% of its

area. Despite its small size, the Arctic Ocean receives 11% of the global riverine discharge [7], introducing freshwater and terrestrial organic material that influence conditions across the entire ocean. In addition, unlike any other ocean, it is almost entirely encircled by land mass, with only limited exchange occurring through shallow and deep water channels. These features make the Arctic Ocean a maritime cul de sac with estuarine-like conditions.

Indifference to temperate and tropical oceans, the Arctic Ocean undergoes immense transformations over an annual cycle. Most notably, the seasonal fluctuations in sea-ice cover. Like the ebb and flow of a tide, sea-ice grows and shrinks on the surface of the Arctic Ocean, from 25% to 75% coverage between summer and winter. Sea-ice is an integral component of the Arctic Ocean. It creates unique habitats, determines water column stratification and influences physicochemical properties of the water below through mediating air-sea gas exchange and by altering light availability [8, 9]. Concurrent with sea ice dynamics, seasonal transformations in the Arctic Ocean are also marked by extreme shifts in light availability, from complete darkness in winter (polar night) to 24 h daylight in summer (polar day). Collectively, these and many other features of the Arctic Ocean make it a remarkably unique and dynamic environment. However, the Arctic Ocean is currently experiencing irreversible perturbations that are redefining its conditions and reshaping the state of ecosystems.

The rate of climate warming in the Arctic is 2.5 times higher than the global average, which has dramatic consequences. The rapid reduction in sea-ice extent and thickness in the Arctic has become one of the global indicators for climate change. Since 1980, winter sea-ice coverage has reduced by 40% and thinned by 1.75 m [10]. As sea-ice continues to disappear, more of the Arctic Ocean is exposed, which in turn absorbs heat and accelerates ice melting, a process known as the ice-albedo effect. It is predicted that ice-free summers may appear as early as 2050 in the Arctic Ocean [11]. Accelerated rates of sea-ice decline in the Eurasian Arctic, have been connected to an increase in heat transport from inflowing Atlantic waters [12]. Atlantic water inflow has doubled in volume in the last 30 years [13] and increased in temperature by 1.4 °C [14], leading to an expansion of Atlantic influence in the Arctic Ocean, termed Atlantification. Through expediting sea-ice decline and increasing subsurface temperatures, Atlantification alters water column stratification and vertical mixing whilst also providing avenues for habitat range expansion of temperate organisms [13, 14]. Notwithstanding other perturbations that are taking place in the Arctic Ocean, these processes are expected to have a profound impact on the ecology and dynamics of microbial communities.

The primary entry point for Atlantic water into the Arctic is the Fram Strait. Located between Greenland and Svalbard, the Fram Strait is the only deep-water gateway into the Arctic and is characterised by two major current systems; the East Greenland Current (EGC), which transports polar water southwards in its upper layer, and the West Spitsbergen Current (WSC) that transports Atlantic water northward (Figure 1). The EGC is responsible for the export of ~50% of freshwater and ~90% of sea-ice from the central Arctic Ocean and thus carries Arctic hydrographic signatures and experiences nearly year-round ice cover [15]. In contrast, the Atlantic water of the WSC maintains a stable temperate of ~5 °C, acting as a heat barrier against sea-ice formation. Atlantic water is also continuously recirculated into the EGC through eddies, although the magnitude varies across latitudes and over temporal scales [16, 17]. The complex and dynamic hydrography of the Fram Strait culminates in distinct water mass regimes of Arctic, mixed and Atlantic origins. This provides an invaluable opportunity to investigate the ecology of microbial communities over a polar to temperate water mass gradient in the context of high-latitude conditions whilst gaining insights into the impact of

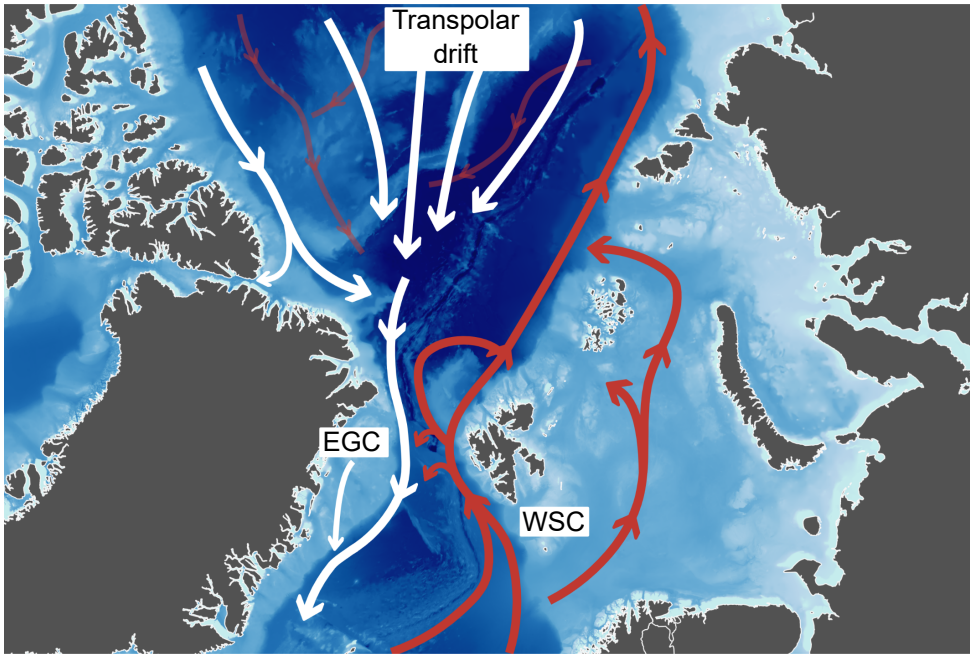


Figure 1. Schematic of the major current features in the Arctic Ocean and Fram Strait. Bathymetry is coloured based on depth (data derived from International Bathymetry Chart of the Arctic Ocean; IBCAO). Arrows indicate currents. Red arrows = Atlantic water, white arrows = polar water. Opaque arrows are surface currents whilst transparent arrows are deep water currents.

Atlantification and future Arctic conditions.

The bulk of the work conducted for this thesis aimed to characterise the ecology of microbial communities across the distinct water mass regimes in the Fram Strait using nucleotide sequencing as the primary investigative tool. Before further describing the research presented in this thesis, a brief background on nucleotide sequencing is necessary as a foundation.

1.2 Nucleotide sequencing

Nucleotide sequencing has revolutionised the field of marine microbial ecology since its introduction in the 1970's. Around the same time that the first sequencing technology was developed, Sanger sequencing [18], work by Woese and Fox revealed the phylogenetic power of the 16S rRNA gene for identifying microbes in the context of evolution [19, 20]. Sequencing of the 16S rRNA gene subsequently became the standard approach for characterising microbial communities in the environment. This led to major developments in our understanding of the phylogenetic diversity of marine microbes as well as their dynamics over space and time. Our capacity to investigate microbial ecology in the environment drastically improved further with the advent of next-generation sequencing (NGS) technologies in the mid 2000's. NGS technologies boasted a throughput >10,000 greater than Sanger sequencing, whilst maintaining low error rates and low cost. Applications of 16S rRNA gene sequencing with NGS platforms unlocked previously unknown diversity in microbial communities, resulting in the discovery of the rare biosphere [21].

The major development that NGS facilitated however, was the capacity to sequence the entire genetic material from a sample in an untargeted approach, termed metagenomics. As a genome encapsulates the genetic content of a single organism, the metagenome is the collection of all genetic content from all genomes of all organisms within a sample. Metagenomics thus provided an unprecedented opportunity to not only investigate the composition of microbes in an environment, but also their functional capacity. Further information about the functions being performed by microbes within the environment was made possible by the development of metatranscriptomics (sequencing of mRNA and rRNA). The meta'omics sequencing era marked a pivotal transition where computational tools became fundamental for the study of microbial ecology. The most significant developments were tools to reconstruct genomes from environmental microbial populations. This involved the stitching together of short NGS reads into longer fragments, known as assembling, and subsequently clustering them based on similar characteristics, a process termed binning. The generated metagenomic bins represent a computational reconstruction of an environmental population genome. To further improve the quality and accuracy of bins, tools were later developed that allowed for interactive viewing and manual curation [22]. This led to the term metagenome-assembled genomes (MAGs), to describe manually curated metagenomic bins. The adoption of metagenomics as a de facto standard tool to investigate microbial communities has resulted in the generation of hundreds of thousands of MAGs and metagenomic bins over the past decade [23, 24], facilitating detailed analyses on diversity, evolution and function. However, such approaches were not without limitations. In particular,

metagenomic assemblies typically only incorporated a proportion of the original reads and failed when faced with genomic regions rich in repeats.

The more recent development of third-generation sequencing technologies (TGS) has provided opportunities to overcome previous computational limitations and gain even deeper insights into microbial communities in the environment. The novel approaches employed by TGS technologies resulted in reads that spanned several genes in length, which would encapsulate repetitive regions and ease the process of assembling genomes. Although the technologies suffered from low throughput and high error rates for some years, the introduction of HiFi reads from PacBio is paving the way for TGS applications in marine microbial ecology. In a single run, modern PacBio platforms can now produce >12 Gbp of HiFi reads with an average length of ~10 kbp and error rates from <0.1 – 1%. The length and quality of HiFi reads offers two key advantages, a) High-quality, long reads can allow for more contiguous assemblies, and b) the encapsulation of several genes within the raw reads facilitates phylogenetic classification and functional annotation of microbial community genetic content without the need for assembling. These platforms are yet to be widely adopted in the field of marine microbial ecology, largely due to their higher cost than NGS platforms. In their current state, it is thus a question of whether the information obtained is worth the cost. As HiFi read metagenomes are at the foundation of the research in this thesis and are the first such datasets from marine microbial communities, the findings presented here can act as a resource for determining the answer to that question.

1.3 Aims of this thesis

The primary aim of this thesis was to characterise the ecology of microbial communities inhabiting high-latitude Arctic and Atlantic waters. More specifically, to characterise the composition and structure of microbial communities, elucidate functional roles, assess dynamics over space and time, investigate the partitioning of ecological niches and identify the main factors that influence microbial ecology at high-latitudes. The work in this thesis was focused on samples collected from different locations in the Fram Strait over multiple years. The Fram Strait harbours water masses of Arctic, mixed and Atlantic origins and thus provided an opportunity to investigate microbial ecology in the context of water mass history and high-latitude conditions. As demonstrated in the research articles, we employed a suite of investigative tools (Figure 2) to address our aim from different fields of view. However, at the foundation of all the research conducted was metagenomics. In particular, this thesis incorporates some of the first applications of long-read metagenomics in marine microbial ecology research. As such, there was a secondary aim to assess the value of using long-read metagenomics to investigate microbial ecology and to develop ways to extract ecological information from such datasets.

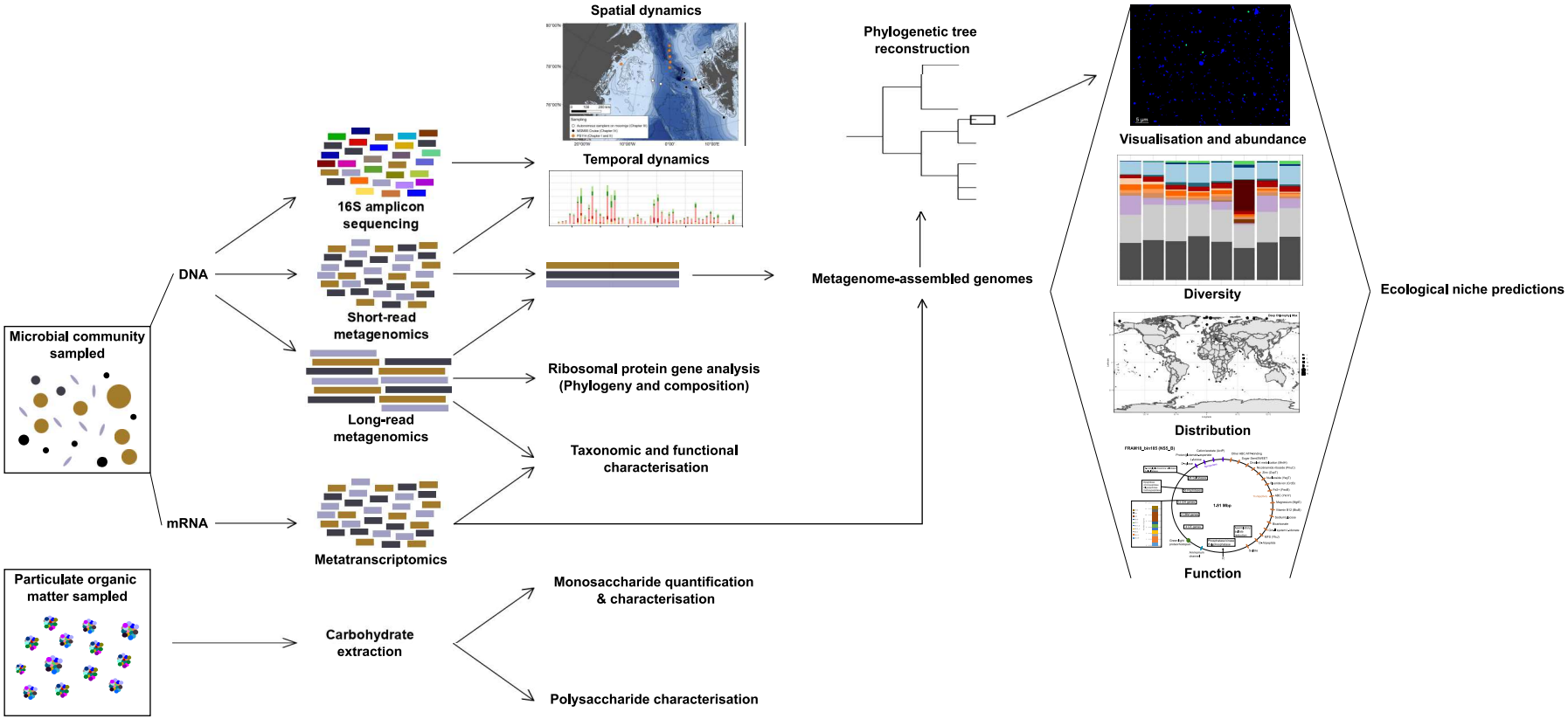


Figure 2. Diagram highlighting the main methods that were applied during this thesis to investigate aspects of microbial ecology in high latitude waters of Arctic, mixed and Atlantic origin.

Chapter II

Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms

Taylor Priest, Luis H. Orellana, Bruno Huettel, Bernhard M. Fuchs and Rudolf Amann

Manuscript published in PeerJ

Contribution of the candidate in % of the total work

Experimental concept and design – 50%

Experimental work/acquisition of experimental data – 10%

Data analysis and interpretation – 90%

Preparation of figures and tables – 100%

Drafting of the manuscript – 80%

PeerJ

Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms

Taylor Priest[1], Luis H. Orellana[1], Bruno Huettel[2], Bernhard M. Fuchs[1] and Rudolf Amann[1]

[1] Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, Bremen, Germany
[2] Max-Planck-Genome-Centre Cologne, Cologne, Germany

ABSTRACT

The impacts of climate change on the Arctic Ocean are manifesting throughout the ecosystem at an unprecedented rate. Of global importance are the impacts on heat and freshwater exchange between the Arctic and North Atlantic Oceans. An expanding Atlantic influence in the Arctic has accelerated sea-ice decline, weakened water column stability and supported the northward shift of temperate species. The only deep-water gateway connecting the Arctic and North Atlantic and thus, fundamental for these exchange processes is the Fram Strait. Previous research in this region is extensive, however, data on the ecology of microbial communities is limited, reflecting the wider bias towards temperate and tropical latitudes. Therefore, we present 14 metagenomes, 11 short-read from Illumina and three long-read from PacBio Sequel II, of the 0.2–3 μm fraction to help alleviate such biases and support future analyses on changing ecological patterns. Additionally, we provide 136 species-representative, manually refined metagenome-assembled genomes which can be used for comparative genomics analyses and addressing questions regarding functionality or distribution of taxa.

Subjects Bioinformatics, Microbiology
Keywords Arctic, Microbiology, Metagenomics, Metagenome-assembled genomes, Microbial ecology

Submitted 8 April 2021
Accepted 14 June 2021
Published 30 June 2021

Corresponding author
Luis H. Orellana,
lorellan@mpi-bremen.de

Academic editor
Thulani Makhalanyane

Additional Information and Declarations can be found on page 13

DOI 10.7717/peerj.11721

INTRODUCTION

The Arctic Ocean is a critical component in the maintenance of Earth's energy balance and the regulation of global climate. Of major importance is the exchange of heat and freshwater between the Arctic and North Atlantic Oceans. The northward transport of Atlantic water is the primary source of heat to the interior of the Arctic Ocean and is vital for water column stability (*Rudels et al., 1994*; *Spall, 2013*). Similarly, the southward transport of Arctic water plays a fundamental role in the thermohaline circulation through the formation of North Atlantic Deepwater (*McGuire et al., 2006*). As is becoming increasingly evident, these processes are experiencing pronounced perturbations as a result of climate change. The inflowing Atlantic water has doubled in volume in the last 30 years (*Oziel et al., 2020*) and increased in temperature by 1.4 °C (*Neukermans, Oziel & Babin, 2018*). This has manifested in an expanding Atlantic presence across the Eurasian Arctic (*Polyakov et al., 2017*), contributing to sea-ice decline (*Lind, Ingvaldsen & Furevik, 2018*),

How to cite this article Priest T, Orellana LH, Huettel B, Fuchs BM, Amann R. 2021. Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms. *PeerJ* **9**:e11721 http://doi.org/10.7717/peerj.11721

warmer subsurface temperatures, a weakening of water column stability and the northward expansion of temperate organisms (*Neukermans, Oziel & Babin, 2018*; *Oziel et al., 2020*). These phenomena are evidence of a much broader and more long-term transition in the state of the Arctic Ocean ecosystem.

The primary region of exchange between the Arctic and North Atlantic Oceans is the Fram Strait. This 450 km wide deep-water gateway, situated between Greenland and Svalbard, is a convergence zone of two distinct hydrographic regimes. The West Spitsbergen Current transports warm and salty Atlantic water northward through the eastern Fram Strait whilst in the western Fram Strait, the East Greenland Current is responsible for 51% of the Arctic Ocean freshwater export (*Serreze et al., 2006*). The convergence of these opposing currents, in a relatively narrow geographical area, provides an invaluable opportunity to investigate the ongoing impacts of climate change on the Arctic Ocean.

Despite decades of research in this region, information regarding the ecology of microbial communities is limited. As primary production increases in the Arctic (*Lewis, Dijken & Arrigo, 2020*) and coastal influences such as thawing permafrost become more pronounced (*Lantuit & Pollard, 2008*; *Vonk et al., 2012*), the availability, quantity and composition of organic matter will change substantially. The primary degraders of organic matter in the marine environment are heterotrophic microbes (*Azam, 1998*). Therefore, characterising their ecology and, in particular, their functional capabilities, may provide insights as to how the Arctic Ocean ecosystem will cope with, and adapt to changing conditions. Currently, a powerful method for addressing such topics is metagenomics.

Widespread efforts in environmental sequencing and the retrieval of metagenome-assembled genomes (MAGs) have been largely directed towards epipelagic communities in temperate and tropical latitudes whilst the polar regions are far less studied. Therefore, we aim to contribute to the alleviation of these biases and present 14 metagenomes (11 from Illumina HiSeq 3000 and three from PacBio Sequel II sequencing platforms) of the 0.2–3 μm fraction from the Fram Strait region. These metagenomic datasets were assembled, binned and manually curated to generate 136 species-representative MAGs.

MATERIALS & METHODS

Sample collection and read generation

Samples ($n = 11$) were collected between July–August of 2018 from the Fram Strait region whilst onboard the RV Polarstern (PS114 cruise). A map of the sampling locations was generated using publically available bathymetric data (*GEBCO Compilation Group, 2020*; *Jakobsson et al., 2020*) and edited using QGIS v3.14.16-Pi (*QGIS.org, 2021*) (Fig. 1). The samples were mostly derived from the deep chlorophyll maximum layer, determined using the in-built Fluorimeter of the Conductivity, Temperature and Depth (CTD)-Rosette sampler during the downcast, and detailed information on their location is provided in Table S1. For each sample, 1 L of seawater was retrieved with a CTD-Rosette sampler and sequentially filtered through a 10 μm, 3 μm and 0.2 μm polycarbonate membrane filter (47 mm diameter) for size fractionation. The full 1 L was filtered through the 10 μm fraction whereas the sample was divided into 2 × 500 ml for the smaller size fractions.

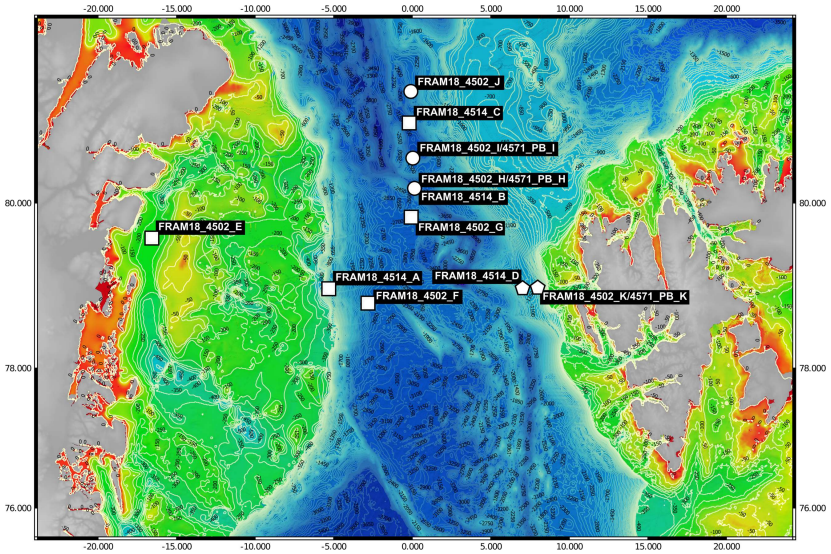


Figure 1 Bathymetric map of sampling locations. The visualised stations were sampled during the PS114 Polarstern cruise in July and August of 2018. The colour scale from red to green indicates depth changes that are more accurately represented by the contour lines. Stations were categorised into three water masses based on temperature and salinity measurements. Squares represent Arctic water mass, pentagons represent Atlantic water mass and circles represent mixed water mass.

Full-size DOI: 10.7717/peerj.11721/fig-1

The filters were immediately frozen and kept at −80 °C until the extraction of DNA. DNA was extracted from one of the 0.2–3 µm fraction filters for each sample (500 ml filtered) following a modified SDS-based extraction method after *Zhou, Bruns & Tiedje (1996)*. The quality of extraction and quantification of DNA was determined using a Qubit 2.0 Fluorometer (Invitrogen, Darmstadt, Germany) (Table 1).

All 11 samples were sequenced at the Max Planck Genome Centre in Cologne, Germany. Sequencing was performed on an Illumina HiSeq 3000 platform, following an ultra-low input library preparation protocol. This resulted in 71–118 million paired-end reads per sample of 150 bp in length (Table 1). Additionally, three of the samples were sequenced on a PacBio Sequel II platform, following an ultra-low library preparation protocol. The ultra-low PacBio library protocol involves a long-range PCR step for AT- and GC-rich sequences followed by a size selection step (removal of sequences < 4.5 kbp). The three samples were barcoded, pooled into a single library and sequenced on a single SMRT Cell. The circular consensus sequencing method was used, generating HiFi reads with >99% per-base accuracy and an output of 4–6 Gbp per sample with an average read length of 8.6–9.6 kbp (Table 1). All sequencing runs were performed without positive or negative controls. To provide an insight into the quality of the obtained sequences, plots of the base quality scores across reads produced by FastqQC (Illumina metagenomes) and NanoPlot

Table 1 **Summary statistics on raw and assembled metagenomes.** Water mass abbreviations are 'Arc = Arctic', 'Mix = mixed', 'Atl = Atlantic'. Library names containing '4514' or '4502' are derived from Illumina HiSeq3000 reads that were assembled with Megahit whilst those containing '4571' are derived from PacBio Sequel II reads that were assembled with MetaFlye.

Sample name	Water mass	DNA yield (ng)	Sequencing platform	Raw read number	Average sequence length (bp)	Estimated coverage	Total assembly length (Mbp)	L50	N50	Number of contigs	Max contig length (Mbp)
FRAM18_4514_A	Arc	25.1	Illumina HiSeq 3000	84,692,914	150	0.79	1,103.249	561	409611	2087649	0.196
FRAM18_4514_B	Mix	530.8	Illumina HiSeq 3000	102,164,804	150	0.86	1,015.001	799	221669	1644515	0.263
FRAM18_4514_C	Arc	65.4	Illumina HiSeq 3000	73,810,275	150	0.72	1,187.025	654	370621	2109170	0.302
FRAM18_4514_D	Atl	613.7	Illumina HiSeq 3000	81,643,335	150	0.73	1,290.853	690	365403	2237087	0.230
FRAM18_4502_E	Arc	243.2	Illumina HiSeq 3000	106,969,873	150	0.8	1,645.447	676	492678	2780604	0.668
FRAM18_4502_F	Arc	231.6	Illumina HiSeq 3000	81,593,071	150	0.74	1,065.963	628	373268	1860335	0.249
FRAM18_4502_G	Arc	189.9	Illumina HiSeq 3000	87,777,577	150	0.76	1,481.554	635	519881	2609315	0.382
FRAM18_4502_H	Arc	277.9	Illumina HiSeq 3000	97,201,780	150	0.82	1,314.445	632	449089	2337881	0.435
FRAM18_4571_H	Arc		Pacbio Sequel II	625,530	9653	0.85	386.515	61541	1295	9700	1.594
FRAM18_4502_I	Mix	303	Illumina HiSeq 3000	71,890,475	150	0.8	961.129	758	231982	1292104	0.409
FRAM18_4571_I	Mix		Pacbio Sequel II	455,246	9665	0.88	317.028	71964	904	7545	2.067
FRAM18_4502_J	Mix	359	Illumina HiSeq 3000	117,925,127	150	0.86	914.466	856	195028	1526032	0.408
FRAM18_4502_K	Atl	190.7	Illumina HiSeq 3000	107,105,259	150	0.83	1,658.809	722	459518	2090439	0.396
FRAM18_4571_K	Atl		Pacbio Sequel II	552,543	8599	0.84	420.259	67691	1226	9944	2.309

Notes.

Water mass abbreviations are 'Arc, Arctic'; 'Mix, mixed; 'Atl, Atlantic;'. Library names containing '4514' or '4502' are derived from Illumina HiSeq3000 reads that were assembled with Megahit whilst those containing '4571' are derived from PacBio Sequel II reads that were assembled with MetaFlye.

v1.32.1 (*De Coster et al., 2018*) (PacBio metagenomes) are provided for a selection of the metagenomes (Fig. S1).

Metagenomic assembly and binning of Illumina reads

Prior to read analysis, Nonpareil v3.3.3 (*Rodriguez & Konstantinidis, 2014*; *Rodriguez et al., 2018*) was used to provide an estimate of the level of coverage of each metagenome (Table 1). As Nonpareil was designed for short reads only, the long PacBio reads were sheared into 150 bp fragments to allow for comparative analysis to the Illumina reads. Coverage values ranged from 0.72–0.88, indicating a high level of coverage with the chosen sequencing depth. Low quality reads and adapters were removed from the Illumina dataset using BBDuk of the BBtools package v38.73 (http://bbtools.jgi.doe.gov/) (parameters: ktrim = r, k = 29, mink = 12, hdist = 1, tbo = t, tpe = t, qtrim = rl, trimq = 20, minlength = 100). Megahit v1.2.9 (*Li et al., 2016*) (parameters: –presets meta-large, –cleaning-rounds 5) was used to assemble short-read metagenomes individually (Table 1).

Quality trimmed reads were mapped to the assemblies using BBMap of the BBtools package (*Bushnell, 2014*) (parameters: minid = 99, idfilter = 97) to provide coverage information for binning. The recovery of MAGs was then performed in a multi-step approach. Firstly, contigs >2.5 kbp in length were binned using three different programs with default settings: Metabat2 v2.12.1 (*Kang et al., 2019*), Concoct v1.1.0 (*Alneberg et al., 2014*) and MaxBin2 v2.2.7 (*Wu, Simmons & Singer, 2016*). A consensus set of non-redundant bins was subsequently retrieved using DasTool v1.1.1 (*Sieber et al., 2018*) and taxonomically classified using CheckM v1.1.2 (*Parks et al., 2015*). In the second step, bins assigned to same taxonomic class were concatenated into a single file and used to recruit raw reads with BBmap (minid = 95, idfilter = 95). The successfully mapped reads from each class were assembled using Megahit (parameters: –presets meta-sensitive –cleaning-rounds 5). The binning pipeline described above was then repeated for each class-level assembly. The taxonomic reassembly was performed as it can greatly improve the quality of MAGs produced through increased contiguity and reduced contamination. The completeness and contamination were determined using CheckM and those that were < 50% complete were removed from further analysis. Reads were recruited to the remaining 218 MAGs using BBmap (parameters: minid = 99 idfilter = 97) to generate coverage information. The MAGs and resulting sequence alignment map files were processed with the metagenomics pipeline implemented in Anvi'o v6.1 (*Eren et al., 2015*). All 218 MAGs were then manually refined using the anvio interactive interface with the anvi-refine function to inspect coverage and reduce contamination where necessary.

Metagenomic assembly and binning of PacBio reads

PacBio HiFi reads were subject to error correction using the program FMLRC v1.0 (*Wang et al., 2018*) with the Illumina quality trimmed reads as a reference. The reads were processed in a similar pipeline as described for the Illumina reads except, metaFlye v2.6 (*Kolmogorov et al., 2020*) (parameters: –pacbio-hifi, -i 5, –genome-size 150 m) was used for assembly (Table 1). To obtain coverage information for binning, the Illumina reads derived from the same sample were used for mapping, using the same parameters described above, due

to the robustness of short-read mapping tools. Taxonomic reassembly was not performed for the PacBio dataset due to the high quality of generated MAGs from single metagenome assemblies. The three PacBio assemblies resulted in 128 consensus MAGs being retrieved after removing those with <50% completeness. All MAGs were then manually refined as described for the Illumina dataset.

The 346 manually refined MAGs were compared using FastANI v1.9 (*Jain et al., 2018*) and grouped into species-level clusters with a genome alignment threshold of 30% and a 95% identity threshold. The highest quality MAG from each cluster (determined by completeness, contamination, N50, number of contigs and the presence of rRNA genes) was designated the representative (Table S2) and used for phylogenetic tree reconstruction. Quality trimmed reads from the Illumina metagenomes were recruited to each species-representative MAG and relative abundances calculated (Table S3). A schematic diagram is provided that summarises the workflow used to analyse the metagenomic data (Fig. 2).

Phylogenetic assessment of MAGs

The taxonomic classification of MAGs was performed using two approaches. Firstly, the classify_wf pipeline of GTDB-tk v1.0.2 (*Chaumeil et al., 2020*; *Parks et al., 2020*) (Release89) was used. Secondly, full-length 16S rRNA gene sequences (>1,400 bp in length) were extracted from representative MAGs with Barrnap v0.9 (*Seeman, 0000*) and placed into the SILVA 138 SSU NR99 reference phylogenetic tree (*Quast et al., 2013*; *Yilmaz et al., 2014*) using the SINA aligner (*Pruesse, Peplies & Glöckner, 2012*) and Maximum Parsimony algorithm of the ARB program (*Ludwig et al., 2004*). In total, 73 out of 136 species-representative MAGs had a complete 16S rRNA gene (>1,400 bp in length) and 23 of those, had more than one gene. The taxonomic assignment of sequences inferred from the SILVA database was used to replace the alpha-numeric taxonomic group names provided by GTDB where possible.

A phylogenetic tree was constructed using a concatenated alignment of 16 ribosomal proteins (L2, L3, L4, L5, L6, L14, L16, L18, L22, L24, S3, S8, S10, S17, S19) following a similar procedure to *Hug et al. (2016)*. To represent the diversity as accurately as possible, the dataset was supplement with genomes of Bacteria and Archaea that were labelled as 'Representative' and 'Complete' in the RefSeq database (>2,500 genomes). Prodigal was used to predict coding sequences and target proteins were identified using hmmsearch v3.3.1 (*Eddy, 2011*) against PFAM HMM models for each ribosomal protein (*E*-value threshold of 1E -5). Individual gene sets were aligned using Muscle v3.8.15 (*Edgar, 2004*) (parameters: -maxiters 16) and trimmed using TrimAI v1.4.1 (*Capella-Gutiérrez, Silla-Martínez & Gabaldón, 2009*) (parameters: -automated1). All alignments were concatenated to form a single 16 gene alignment and a phylogenetic tree constructed using FastTree v.2.1.10 (*Price, Dehal & Arkin, 2010*) (parameters: -gamma –lg). The tree was visualised and annotated using iToL v4 (*Letunic & Bork, 2019*) (Fig. 3).

Recovery of full-length 16S rRNA gene sequences

The extraction of 16S rRNA gene sequences from metagenomic reads can provide an insight into the community sampled and aid in identifying major taxonomic groups that are missed

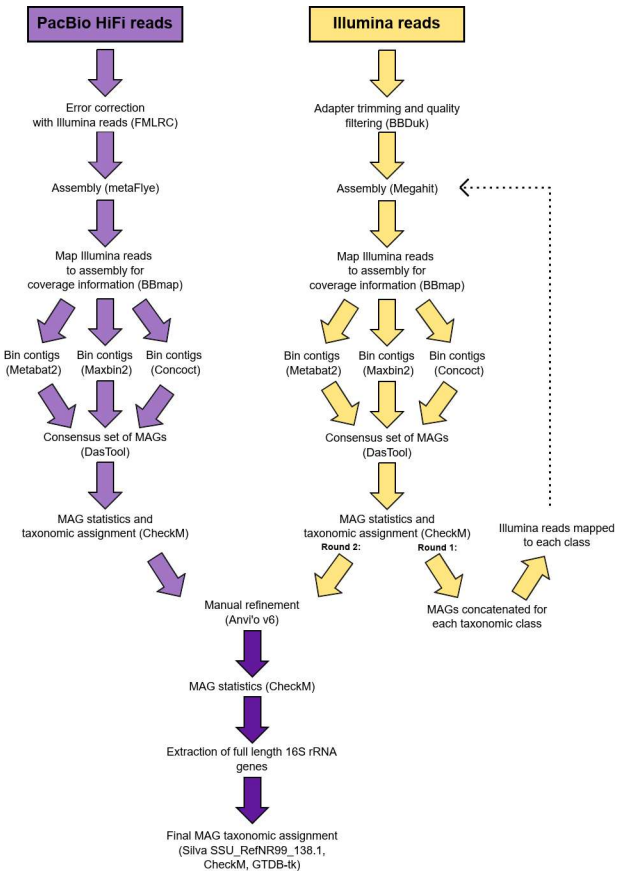


Figure 2 **Schematic diagram of the workflow used to process the metagenomic data.**
Full-size DOI: 10.7717/peerj.11721/fig-2

in the recovery of MAGs. Additionally in this study, it also allows for the comparison of the two sequencing methods. However, a major restriction with short-read metagenomic sequencing is the limited capacity to accurately reassemble full length 16S rRNA genes. With the advent of highly accurate long read sequences generated from PacBio sequel II (>99% accuracy), full length 16S rRNA genes can be retrieved from single reads without a need for assembly, thus circumventing previous limitations. This not only results in MAGs with complete 16S rRNA operons but also provides many additional sequences that can help to expand current databases without the need for a second, targeted sequencing run or additional sequencing platform. To provide future users of the data with a more detailed insight into the community and to demonstrate the value of long PacBio reads, we used the tool Barrnap v0.9 with default settings (parameters: –kingdom bac) to extract

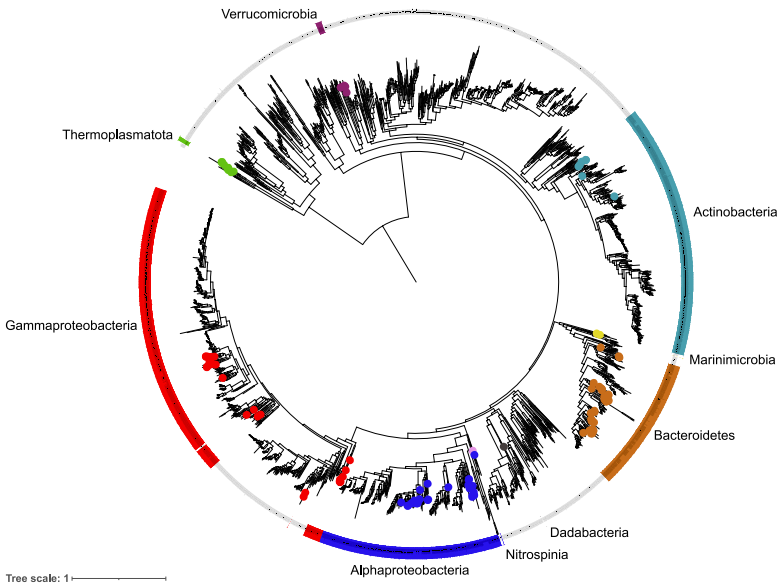


Figure 3 Phylogenetic diversity of metagenome assembled genomes (MAGs) from the Fram Strait. The maximum likelihood tree was constructed from the concatenated alignments of 16 ribosomal proteins present in the MAGs and reference genomes of Bacteria and Archaea available in RefSeq. Coloured outer rings indicate groups that are represented by MAGs whilst circles on midpoints of the same colour indicate the exact position of MAGs within those groups.

Full-size DOI: 10.7717/peerj.11721/fig-3

16S rRNA gene sequences from the PacBio reads. A length cut-off was applied at 1000 bp, to focus on complete or near-complete gene sequences. For comparison, 16S rRNA reads were identified in the Illumina dataset using SortMeRna (*Kopylova, Noé & Touzet, 2012*) with a length cut-off of 120 bp. The extracted reads from each dataset were clustered into operational taxonomic units (OTUs) using the program CD-HIT-EST (*Fu et al., 2012*) at a 99% threshold. Reads were aligned using the SINA aligner and phylogenetically placed into the SILVA SSU_RefNR99 138.1 reference tree using the Parsimony tool in ARB, as described in '*Phylogenetic assessment of MAGs*'. The raw read numbers for the identified community were Hellinger transformed, compared using a Bray–Curtis dissimilarity matrix and visualised in a dendrogram format and ordinated using non-metric multi-dimensional scaling analysis (NMDS; Fig. S2) using the vegan package (*Oksanen et al., 2013*) in RStudio v1.1.463 (*R Core Team, 2015*). The taxonomic diversity of each sample was visualised using the ggplot2 package (*Wickham, 2016*) in RStudio v1.1.463 (Fig. 4) and the relative abundance of all taxonomic groups across samples is provided in Tables S4 and S5 (Illumina-derived 16S rRNA gene composition) and Table S4 (PacBio-derived 16S rRNA gene composition).

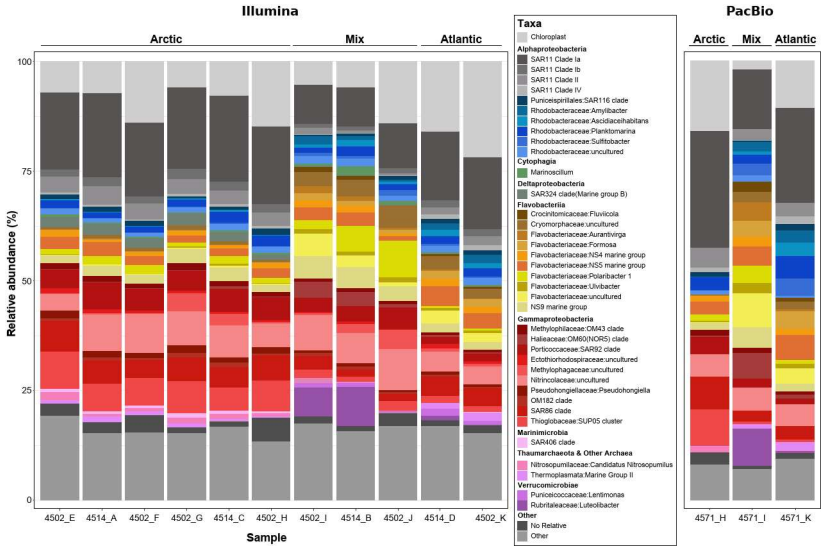


Figure 4 Phylogenetic composition of each metagenome sample derived from 16S rRNA gene sequences. Genes were recovered from raw metagenomic reads using Barrnap (PacBio reads) and SortMeRNA (Illumina reads) and included only if the length was >1,000 bp or >120 bp respectively. Sequences from each dataset were clustered at a 99% identity threshold and taxonomically classified by inclusion into a reference phylogenetic tree from the SILVA 138 NR99 database. For clarity, the lettering used at the end of the sample name shows which PacBio and Illumina metagenomes came from the same sample, e.g., 4571_H and 4502_H.

Full-size DOI: 10.7717/peerj.11721/fig-4

RESULTS

Identification of distinct water masses

The samples were identified as being of either Atlantic, Arctic or a mixed water mass origin based upon the measured abiotic parameters (Table S1). Those with a temperature < 0 °C and a salinity of < 34 psu were labelled as Arctic whereas those with a temperature of >5 °C and a salinity of ~35 psu were labelled as Atlantic, in accordance with previous studies (*Rudels et al., 2013*; *Fadeev et al., 2018*). Samples with values in between these thresholds were defined as being of mixed origin.

Coverage and community

To determine the quality of the 14 metagenomes with respect to capturing the sampled community, NonPareil was used to calculate coverage; for this analysis, the long PacBio reads were sheared into 150 bp fragments to allow a direct comparison to the short Illumina reads. The coverage values across all metagenomes ranged from 0.72–0.88, indicating a high coverage regardless of sequencing method.

Prior to assembling the metagenomic reads, the community composition of each sample was assessed using the 16S rRNA gene (Fig. 4). Such an analysis provides an insight into

the community sampled for future users of the data, whilst allowing comparisons between the two sequencing methods and between the sampled community and the phylogenetic diversity of MAGs recovered. In total, >950,000 16S rRNA gene fragments were recovered from short-read metagenomes along with >50,000 from long-read metagenomes. Within those recovered from long-reads, >16,000 were identified as full-length, highlighting the value of PacBio Sequel II sequences. Comparing the community composition identified across samples resulted in three distinct clusters reflecting the Artic, Atlantic and mixed water masses (Fig. S2) that were identified based on variations in abiotic parameters (Table S1).

Assembly

The eleven short-read metagenomes, containing between 46–81 million reads post quality trimming, were individually assembled using Megahit, generating assemblies with an average total length of 1.05 Gbp, N50 value of 691 and 2.1 million contigs (Table 1). The three long-read metagenomes were individually assembled using metaFlye which resulted, on average, in shorter assemblies than the short-read metagenomes, 374 Mbp in length, but with much higher N50 values, 67 kbp, and a lower number of contigs, 9063 (Table 1). The average longest contig length was also significantly larger in the long-read assemblies, 1.99 Mbp, compared to the short-read assemblies, 0.34 Mbp.

Binning

The contigs longer than 2.5 kbp were used to recruit reads from all short-read Illumina metagenomes using BBmap, to provide as much coverage variation information as possible to the binning tools. After binning with three different tools, DasTool recovered a consensus set of 349 bins that were >50% complete and < 10% contaminated. All bins were manually refined using anvi'o before being dereplicated into 136 species-level clusters, based on a 95% ANI threshold (Table S2). The selected species-representative MAGs had completeness values between 52–100% and between 0–9% contamination with an average genome size of 1.94 Mbp (Table S2). Of these, 27 species-representative MAGs were classified as high-quality drafts according to the MIMAGs standards(*Bowers et al., 2017*); meaning they contain a 23S, 16S and 5S rRNA gene and at least 18 tRNAs with a completeness value of >90% and contamination < 5%. Furthermore, owing to the long PacBio reads, 75 of the MAGs had a 16S rRNA gene present and 35 of the MAGs were composed of < 10 contigs, indicating a very high contiguity.

The phylogenetic classification of MAGs was performed using two different approaches to ensure robustness and reliability, these included a 16S rRNA gene approach, where the genes were present, and a single copy marker gene and phylogenetic approach through the GTDBtk tool. The diversity captured by species-representative MAGs was then visualised through the reconstruction of a phylogenetic tree using a concatenated alignment of 16 ribosomal proteins from the MAGs and >2500 genomes labelled as 'Complete' and 'Representative' in the RefSeq database (Fig. 3). The recovered diversity encompassed 9 phyla, 11 classes, 27 orders, ~51 families and ~54 genera. The most species-rich taxa were the Flavobacteriales (41 species), Pseudomonadales (18 species) and Rhodobacterales (17

species). This picture of community diversity obtained from the MAGs is comparable to that from the 16S rRNA gene sequences alone, even though we are able to recover a much higher number of 16S rRNA gene sequences.

Relative abundance of MAGs

The range in estimated relative abundance values across the MAGs were from < 0.001–10.75%, with the lower values being attributed to the recruitment of reads to MAGs from compositionally different metagenomes (Tables S4 and S5). Summing the relative abundance values in each sample indicated that the species-representative MAGs accounted for 52.3–89.7% of the community, further supporting that the recovered MAGs covered all of the major taxonomic groups present.

DISCUSSION

The Fram Strait is not only a region of global significance due to its role in heat and water mass exchange but also as it provides an invaluable opportunity to study ecological changes from the Atlantic to Arctic Ocean. Although this region has been studied extensively in recent years, there is still only limited information available on the ecology of microbial communities. The metagenomics and MAG dataset presented here is derived from samples collected across the Fram Strait region and provides unique genetic resources represented in contrasting water masses of Arctic, Atlantic and mixed origin. The dataset also provides a valuable combination of short-read and long-read metagenomes, representing one of the first PacBio Sequel II metagenome and MAG dataset from marine environmental samples.

The distinct water masses sampled across the Fram Strait are distinguishable based on temperature and salinity (*Rudels et al., 2013*) and are shown here to harbor unique microbial community compositions (Fig. 4 and Fig. S2). One major distinction is the elevated proportions of Flavobacteria taxa (such as Aurantivirga, Formosa and NS5) in the Atlantic (West Spitsbergen Current; WSC) compared to the Arctic (East Greenland Current; EGC) water mass, which is likely influenced by the time of sampling (July–August). Seasonal phytoplankton blooms in the WSC region have been well evidenced and shown to reach maximum integrated chlorophyll *a* values of 100 mg/m^3 (*Nöthig et al., 2015*). Summer phytoplankton blooms typically occur from June to July (*Nöthig et al., 2015*) and lead to the enrichment of Flavobacteria, with intermittent peaks of specific taxa (Formosa, Polaribacter and NS5) (*Wietz et al., 2021*) resembling successional patterns that are known from temperate spring phytoplankton blooms (*Teeling et al., 2012*). In comparison, the EGC does not experience such pronounced phytoplankton blooms and instead it has been suggested that a different food web-based structure may exist in these waters (*Wietz et al., 2021*). In agreement with previous findings (*Fadeev et al., 2018*), the EGC was enriched in Gammaproteobacteria (SAR86, SUP05) and taxa related to Arctic winter and deeper waters (Marinimicrobia and SAR324). Between these distinct water masses, the central Fram Strait region is subject to complex and dynamic hydrographic processes with lateral mixing, advection of Atlantic water under Arctic water and westward flowing-mesoscale eddies originating from the WSC all exerting an influence over different spatial and temporal scales. Due to such complexities in determining these features, we defined the

samples from the central Fram Strait region, whose abiotic parameters were between the thresholds of Arctic and Atlantic water, as 'mixed'. These mixed origin samples were shown to harbor the highest proportion of taxa within the Cryomorphaceae and Flavobacteriaceae (such as Polaribacter and NS9) as well as consist of up to 10% Verrucomicrobia that was in low (<2%) abundance in the other water masses. Recently, an investigation into a mesoscale filament in the central Fram Strait region revealed an increase in phytoplankton productivity, microbial cell counts and specific taxa related to phytoplankton-derived organic matter degradation (*Fadeev et al., 2021*). Although in the data presented here, the measured fluorescence values (Table S2) were not indicative of a bloom event, the enriched taxa are known as key players in organic matter degradation. Therefore, it is possible that the mixed samples are derived from a mesoscale filament or eddy and represent a post-phytoplankton bloom situation.

Integrating short and long read sequencing technologies to recover microbial populations

The pipeline we employed to process metagenomics reads was carefully optimized to ensure high quality and accurate assemblies and to maximize the number of near-complete MAGs recovered. The assembly of reads for each metagenome individually as opposed to using a co-assembly approach, likely reduces the chance of chimera formation and prevents the loss of strain variation across populations. The subsequent binning of contigs, after removing those less than 2.5 kbp in length to minimize misbinning (*Chen et al., 2020*), was performed using multiple tools as opposed to a single tool as this approach has recently been shown to greatly increase the number of reconstructed near-complete genomes (*Probst et al., 2017*; *Sieber et al., 2018*). To further improve the completeness and contiguity of bins, we reassembled reads that were recruited to bins of the same taxonomic class within each sample. Although the resulting set of bins were of seemingly high quality, it is well known that using automated tools can result in misbinning of contigs due to similarities in sequence composition and coverage across genomic regions of different microbial populations (*Chen et al., 2020*). Therefore, each of the generated bins was visually inspected and subject to manual refinement which involved the removal of misplaced contigs and the discarding of erroneous bins.

By employing long-read and short-read sequencing, we are able to compare the number and quality of MAGs retrieved between both platforms. Of the species-representative MAGs recovered, those generated from the PacBio metagenomes had, on average, larger genome sizes, higher N50 values and were less fragmented compared to those retrieved from Illumina metagenomes (Table S2). One of the major limitations of short-read metagenomics is the low recovery rate of rRNA genes within MAGs. However, in this study, 84% of MAGs retrieved from the PacBio metagenomes contained at least one complete 16S rRNA gene sequence, highlighting another key advantage of using long Hifi reads. Therefore, we can conclude that HiFi read metagenomes derived from the PacBio Sequel II platform can greatly improve the number and quality of MAGs recovered, which will allow for further advancement in our understanding of the ecology of marine microbial communities.

CONCLUSION

The aim of this manuscript was to provide a useful data resource to supplement future ecological analyses on Arctic microbial communities and to help alleviate biases against metagenomic sequence data from polar regions. The generation of 14 metagenomes from short and long read sequencing platforms along with 136 manually-refined species-representative MAGs provides a valuable dataset to address questions regarding distribution of taxa and functionality on a community- and species-level as well for downstream comparative genomics.

An initial insight into the composition of the metagenomes using 16S rRNA gene sequences revealed taxonomically-rich communities with distinct compositions corresponding to the different water masses sampled. The recovery of more than 16,000 full-length 16S rRNA gene sequences from raw PacBio reads can allow for further high-resolution phylogenetic analyses to be performed. The diversity captured by the 136 manually-refined species-representative MAGs encompassed more than 50 genera and consisted of members from all major taxonomic groups in the sampled community. Furthermore, the majority of MAGs recovered were of high quality, with 27 MAGs being classified as high-quality drafts according to MIMAGS standards, 75 MAGs containing at least one 16S rRNA gene and 35 MAGs having < 10 contigs.

The pipeline used to process the metagenome data and recover the described MAGs was thoroughly tested and optimized at each stage to ensure reliable and high-quality results. Although the provided data is suitable for direct inclusion in further analyses, it is recommended to confirm any of the stated values here using the most up to date analysis tools, particularly with respect to MAG completeness, contamination and taxonomic classification.

DATA RECORDS

All data provided in this study has been deposited in the European Nucleotide Archive (ENA) at EMBL-EBI under accession number PRJEB41592.

ACKNOWLEDGEMENTS

We would like to thank Jörg Wulf and Mirja Meiners, from the Molecular Ecology department at the Max Planck Institute for Marine Microbiology in Bremen, for their technical support. Furthermore we would like to thank the team at the Max Planck Genome Centre in Cologne for their efforts with sequencing the samples.

ADDITIONAL INFORMATION AND DECLARATIONS

Funding

This work was funded by the Max Planck Society. The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Grant Disclosures

The following grant information was disclosed by the authors:
The Max Planck Society.

Competing Interests

The authors declare there are no competing interests.

Author Contributions

- Taylor Priest conceived and designed the experiments, performed the experiments, analyzed the data, prepared figures and/or tables, authored or reviewed drafts of the paper, and approved the final draft.
- Luis H. Orellana, Bernhard M. Fuchs and Rudolf Amann conceived and designed the experiments, authored or reviewed drafts of the paper, and approved the final draft.
- Bruno Huettel performed the experiments, authored or reviewed drafts of the paper, and approved the final draft.

Data Availability

The following information was supplied regarding data availability:

All data in this study are available in the European Nucleotide Archive (ENA) at EMBL-EBI: PRJEB41592.

Supplemental Information

Supplemental information for this article can be found online at http://dx.doi.org/10.7717/peerj.11721#supplemental-information.

REFERENCES

Alneberg J, Bjarnason BS, Bruijn IDe, Schirmer M, Quick J, Ijaz UZ, Lahti L, Loman NJ, Andersson AF, Quince C. 2014. Binning metagenomic contigs by coverage and composition. Nature. *Nature Methods* **11**:1144–1146 DOI 10.1038/nmeth.3103.

Azam F. 1998. Microbial control of oceanic carbon flux: the plot thickens. *Science* **280**:694–696 DOI 10.1126/science.280.5364.694.

Bowers RM, Kyrpides NC, Stepanauskas R, Harmon-Smith M, Doud D, Reddy TBK, Schulz F, Jarett J, Rivers AR, Eloe-Fadrosh EA, Tringe SG, Ivanova NN, Copeland A, Clum A, Becraft ED, Malmstrom RR, Birren B, Podar M, Bork P, Weinstock GM, Garrity GM, Dodsworth JA, Yooseph S, Sutton G, Glöckner FO, Gilbert JA, Nelson WC, Hallam SJ, Jungbluth SP, Ettema TJG, Tighe S, Konstantinidis KT, Liu W-T, Baker BJ, Rattei T, Eisen JA, Hedlund B, McMahon KD, Fierer N, Knight R, Finn R, Cochrane G, Karsch-Mizrachi I, Tyson GW, Rinke C, Lapidus A, Meyer F, Yilmaz P, Parks DH, Eren AM, Schriml L, Banfield JF, Hugenholtz P, Woyke T. 2017. Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology* **35**:725–731 DOI 10.1038/nbt.3893.

Bushnell B. 2014. BBTools software package. *http://bbtools.jgi.doe.gov*.

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973 DOI 10.1093/bioinformatics/btp348.

Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. 2020. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *Bioinformatics* **36**:1925–1927 DOI 10.1093/bioinformatics/btz848.

Chen L-X, Anantharaman K, Shaiber A, Eren AM, Banfield JF. 2020. Accurate and complete genomes from metagenomes. *Genome Research* **30**:315–333 DOI 10.1101/gr.258640.119.

De Coster W, D'Hert S, Schultz DT, Cruts M, Van Broeckhoven C. 2018. NanoPack: visualizing and processing long-read sequencing data. *Bioinformatics* **34**:2666–2669 DOI 10.1093/bioinformatics/bty149.

Eddy SR. 2011. Accelerated profile HMM searches. *PLOS Computational Biology* **7**:e1002195 DOI 10.1371/journal.pcbi.1002195.

Edgar RC. 2004. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Research* **32**:1792–1797 DOI 10.1093/nar/gkh340.

Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, Delmont TO. 2015. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ* **3**:e1319 DOI 10.7717/peerj.1319.

Fadeev E, Salter I, Schourup-Kristensen V, Nöthig E-M, Metfies K, Engel A, Piontek J, Boetius A, Bienhold C. 2018. Microbial communities in the east and west fram strait during sea ice melting season. *Frontiers in Marine* **5**:429 DOI 10.3389/fmars.2018.00429.

Fadeev E, Wietz M, Appen W-J von, Iversen MH, Nöthig E-M, Engel A, Grosse J, Graeve M, Boetius A. 2021. Submesoscale physicochemical dynamics directly shape bacterioplankton community structure in space and time. *Limnology and Oceanography* Epub ahead of print May 26 2021 DOI 10.1002/lno.11799.

Fu L, Niu B, Zhu Z, Wu S, Li W. 2012. CD-HIT: accelerated for clustering the next-generation sequencing data. *Bioinformatics (Oxford, England)* **28**:3150–3152 DOI 10.1093/bioinformatics/bts565.

GEBCO Compilation Group. 2020. GEBCO 2020 Grid.

Hug LA, Baker BJ, Anantharaman K, Brown CT, Probst AJ, Castelle CJ, Butterfield CN, Hernsdorf AW, Amano Y, Ise K, Suzuki Y, Dudek N, Relman DA, Finstad KM, Amundson R, Thomas BC, Banfield JF. 2016. A new view of the tree of life. *Nature Microbiology* **1**:1–6 DOI 10.1038/nmicrobiol.2016.48.

Jain C, Rodriguez RLM, Phillippy AM, Konstantinidis KT, Aluru S. 2018. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nature Communications* **9**:5114 DOI 10.1038/s41467-018-07641-9..

Jakobsson M, Mayer LA, Bringensparr C, Castro CF, Mohammad R, Johnson P, Ketter T, Accettella D, Amblas D, An L, Arndt JE, Canals M, Casamor JL, Chauché N, Coakley B, Danielson S, Demarte M, Dickson M-L, Dorschel B, Dowdeswell JA, Dreutter S, Fremand AC, Gallant D, Hall JK, Hehemann L, Hodnesdal H, Hong J,

Ivaldi R, Kane E, Klaucke I, Krawczyk DW, Kristoffersen Y, Kuipers BR, Millan R, Masetti G, Morlighem M, Noormets R, Prescott MM, Rebesco M, Rignot E, Semiletov I, Tate AJ, Travaglini P, Velicogna I, Weatherall P, Weinrebe W, Willis JK, Wood M, Zarayskaya Y, Zhang T, Zimmermann M, Zinglersen KB. 2020. The International Bathymetric Chart of the Arctic Ocean Version 4.0. *Scientific Data* **7**:176 DOI 10.1038/s41597-020-0520-9.

Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, Wang Z. 2019. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ* **7**:e7359 DOI 10.7717/peerj.7359.

Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, Kuhn K, Yuan J, Polevikov E, Smith TPL, Pevzner PA. 2020. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nature Methods* **17**:1103–1110 DOI 10.1038/s41592-020-00971-x.

Kopylova E, Noé L, Touzet H. 2012. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *Bioinformatics* **28**:3211–3217 DOI 10.1093/bioinformatics/bts611.

Lantuit H, Pollard WH. 2008. Fifty years of coastal erosion and retrogressive thaw slump activity on Herschel Island, southern Beaufort Sea, Yukon Territory, Canada. *Geomorphology* **95**:84–102 DOI 10.1016/j.geomorph.2006.07.040.

Letunic I, Bork P. 2019. Interactive tree of life (iTOL) v4: recent updates and new developments. *Nucleic Acids Research* **47**:W256–W259 DOI 10.1093/nar/gkz239.

Lewis KM, Dijken GL van, Arrigo KR. 2020. Changes in phytoplankton concentration now drive increased Arctic Ocean primary production. *Science* **369**:198–202 DOI 10.1126/science.aay8380.

Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, Yamashita H, Lam T-W. 2016. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. *Methods* **102**:3–11 DOI 10.1016/j.ymeth.2016.02.020.

Lind S, Ingvaldsen RB, Furevik T. 2018. Arctic warming hotspot in the northern Barents Sea linked to declining sea-ice import. *Nature Climate Change* **8**:634–639 DOI 10.1038/s41558-018-0205-y.

Ludwig W, Strunk O, Westram R, Richter L, Meier H, null Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüssmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer K-H. 2004. ARB: a software environment for sequence data. *Nucleic Acids Research* **32**:1363–1371 DOI 10.1093/nar/gkh293.

McGuire AD, Chapin FS, Walsh JE, Wirth C. 2006. Integrated regional changes in arctic climate feedbacks: implications for the global climate system. *Annual Review of Environment and Resources* **31**:61–91 DOI 10.1146/annurev.energy.31.020105.100253.

Neukermans G, Oziel L, Babin M. 2018. Increased intrusion of warming Atlantic water leads to rapid expansion of temperate phytoplankton in the Arctic. *Global Change Biology* **24**:2545–2553 DOI 10.1111/gcb.14075.

Nöthig E-M, Bracher A, Engel A, Metfies K, Niehoff B, Peeken I, Bauerfeind E, Cherkasheva A, Gäbler-Schwarz S, Hardge K, Kilias E, Kraft A, Kidane YM, Lalande C, Piontek J, Thomisch K, Wurst M. 2015. Summertime plankton ecology in Fram Strait—a compilation of long- and short-term observations. *Polar Research* **34**:23349 DOI 10.3402/polar.v34.23349.

Oksanen J, Blanchet FG, Kindt R, Legendre P, Minchin P, O'Hara R, Simpson G, Solymos P, Stevens M, Wagner H. 2013. Vegan: community Ecology Package. R Package Version. 2.0-10. CRAN.

Oziel L, Baudena A, Ardyna M, Massicotte P, Randelhoff A, Sallée J-B, Ingvaldsen RB, Devred E, Babin M. 2020. Faster Atlantic currents drive poleward expansion of temperate phytoplankton in the Arctic Ocean. *Nature Communications* **11**:1–8 DOI 10.1038/s41467-020-15485-5.

Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. 2020. A complete domain-to-species taxonomy for Bacteria and Archaea. *Nature Biotechnology* **38**:1079–1086 DOI 10.1038/s41587-020-0501-8.

Parks DH, Imelfort M, Skennerton CT, Hugenholtz P, Tyson GW. 2015. CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research* **25**:1043–1055 DOI 10.1101/gr.186072.114.

Polyakov IV, Pnyushkov AV, Alkire MB, Ashik IM, Baumann TM, Carmack EC, Goszczko I, Guthrie J, Ivanov VV, Kanzow T, Krishfield R, Kwok R, Sundfjord A, Morison J, Rember R, Yulin A. 2017. Greater role for Atlantic inflows on sea-ice loss in the Eurasian Basin of the Arctic Ocean. *Science* **356**:285–291 DOI 10.1126/science.aai8204.

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2 –approximately maximum-likelihood trees for large alignments. *PLOS ONE* **5**:e9490 DOI 10.1371/journal.pone.0009490.

Probst AJ, Castelle CJ, Singh A, Brown CT, Anantharaman K, Sharon I, Hug LA, Burstein D, Emerson JB, Thomas BC, Banfield JF. 2017. Genomic resolution of a cold subsurface aquifer community provides metabolic insights for novel microbes adapted to high CO2 concentrations. *Environmental Microbiology* **19**:459–474 DOI 10.1111/1462-2920.13362.

Pruesse E, Peplies J, Glöckner FO. 2012. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics (Oxford, England)* **28**:1823–1829 DOI 10.1093/bioinformatics/bts252.

QGIS.org. 2021. QGIS Association. *Available at https://www.qgis.org*.

Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, Peplies J, Glöckner FO. 2013. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research* **41**:D590–D596 DOI 10.1093/nar/gks1219.

R Core Team. 2015. A language and environment for statistical computing. Vienna, Austria: R Foundation for statistical computing.

Rodriguez RLM, Gunturu S, Tiedje JM, Cole JR, Konstantinidis KT. 2018. Nonpareil 3: Fast Estimation of Metagenomic Coverage and Sequence Diversity. *MSystems* **3**:e00039–18 DOI 10.1128/mSystems.00039-18.

Rodriguez RLM, Konstantinidis KT. 2014. Nonpareil: a redundancy-based approach to assess the level of coverage in metagenomic datasets. *Bioinformatics* **30**:629–635 DOI 10.1093/bioinformatics/btt584.

Rudels B, Jones EP, Anderson LG, Kattner G. 1994. On the intermediate depth waters of the Arctic Ocean. In: Johannessen OM, Muench RD, Overland JE, eds. *The role of the Polar Oceans in Shaping the Global Climate*. Washington, D.C.: American Geophysical Union, 33–46 DOI 10.1029/GM085p0033.

Rudels B, Schauer U, Björk G, Korhonen M, Pisarev S, Rabe B, Wisotzki A. 2013. Observations of water masses and circulation with focus on the Eurasian Basin of the Arctic Ocean from the 1990s to the late 2000s. *Science* **9**:147–169 DOI 10.5194/os-9-147-2013.

Seeman T. barrnap 0.9 0.9: rapid ribosomal RNA prediction. *Available at https://github.com/tseeman/barrnap*.

Serreze MC, Barrett AP, Slater AG, Woodgate RA, Aagaard K, Lammers RB, Steele M, Moritz R, Meredith M, Lee CM. 2006. The large-scale freshwater cycle of the Arctic. *Journal of Geophysical Research: Oceans* **111**:C11010 DOI 10.1029/2005JC003424.

Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, Banfield JF. 2018. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nature Microbiology* **3**:836–843 DOI 10.1038/s41564-018-0171-1.

Spall MA. 2013. On the Circulation of Atlantic Water in the. *Journal of Physical Oceanography* **43**:2352–2371 DOI 10.1175/JPO-D-13-079.1.

Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM, Kassabgy M, Huang S, Mann AJ, Waldmann J, Weber M, Klindworth A, Otto A, Lange J, Bernhardt J, Reinsch C, Hecker M, Peplies J, Bockelmann FD, Callies U, Gerdts G, Wichels A, Wiltshire KH, Glöckner FO, Schweder T, Amann R. 2012. Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science* **336**:608–611 DOI 10.1126/science.1218344.

Vonk JE, Sánchez-García L, Van Dongen BE, Alling V, Kosmach D, Charkin A, Semiletov IP, Dudarev OV, Shakhova N, Roos P, Eglinton TI, Andersson A, Ö Gustafsson. 2012. Activation of old carbon by erosion of coastal and subsea permafrost in Arctic Siberia. *Nature* **489**:137–140 DOI 10.1038/nature11392.

Wang JR, Holt J, McMillan L, Jones CD. 2018. FMLRC: hybrid long read error correction using an FM-index. *BMC Bioinformatics* **19**:50 DOI 10.1186/s12859-018-2051-3.

Wickham H. 2016. ggplot2: elegant graphics for data analysis. Springer-Verlag New York.

Wietz M, Bienhold C, Metfies K, Torres-Valdés S, von AppenW-J, Salter I, Boetius A. 2021. The polar night shift: annual dynamics and drivers of microbial community structure in the Arctic Ocean. ArXiv preprint. arXiv:2021.04.08.436999.

Wu Y-W, Simmons BA, Singer SW. 2016. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. *Bioinformatics* **32**:605–607 DOI 10.1093/bioinformatics/btv638.

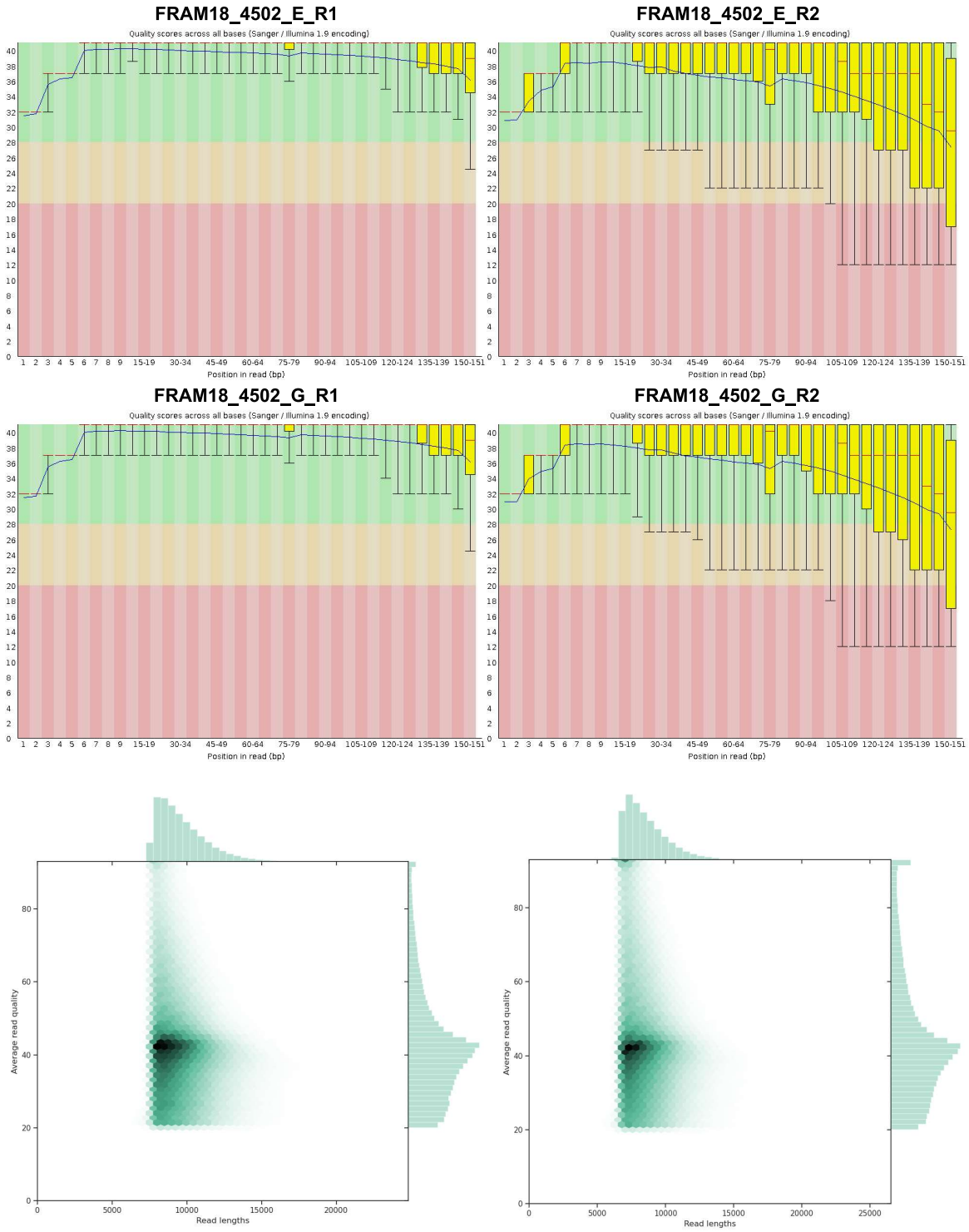
Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, Schweer T, Peplies J, Ludwig W, Glöckner FO. 2014. The SILVA and All-species Living Tree Project (LTP) taxonomic frameworks. *Nucleic Acids Research* **42**:D643–D648 DOI 10.1093/nar/gkt1209.

Zhou J, Bruns MA, Tiedje JM. 1996. DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology* **62**:316–322 DOI 10.1128/aem.62.2.316-322.1996.

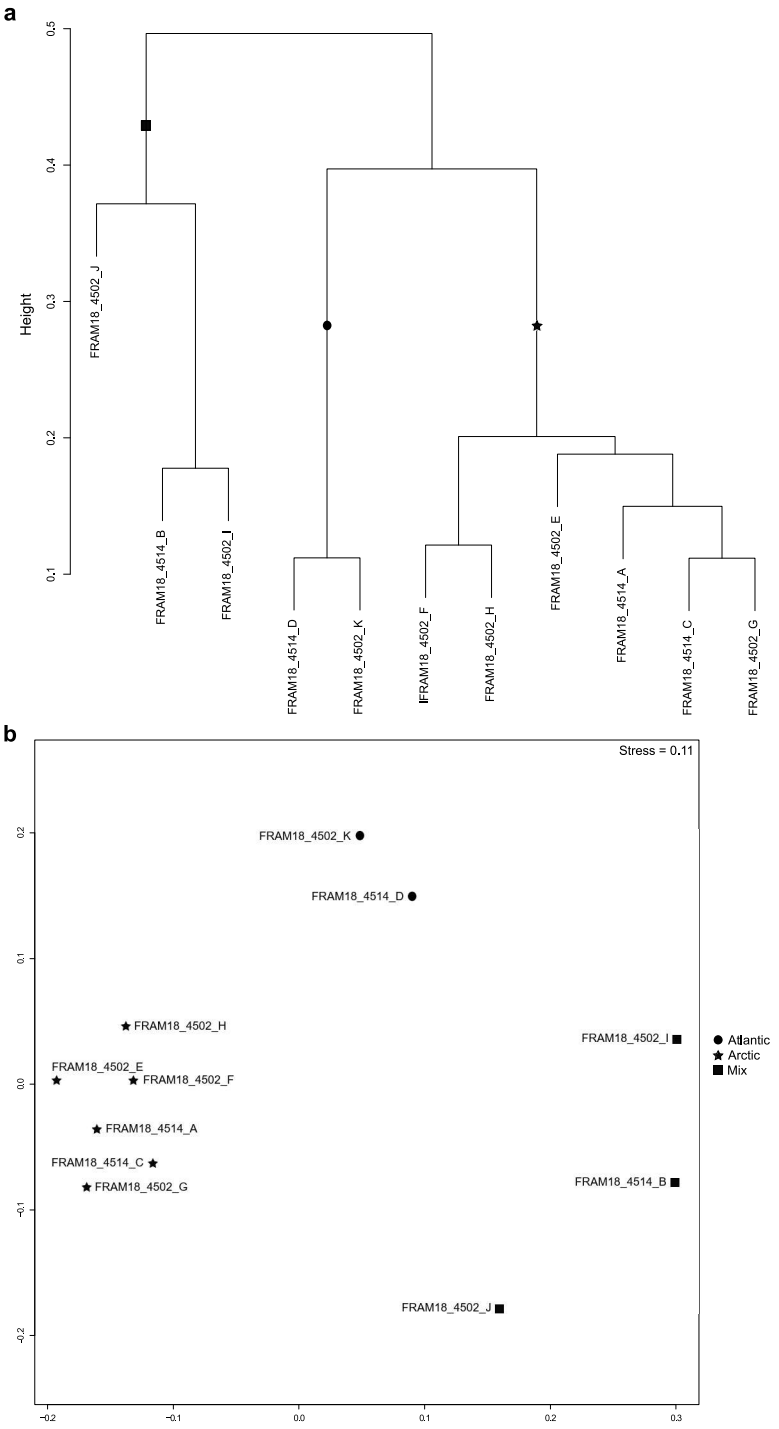
Supplementary tables and figures

All supplementary tables are available on the USB attached with this thesis and at the following link. Due to their large size, they were not included in the printed thesis.

https://owncloud.mpi-bremen.de/index.php/s/ZuO907DWy8Vu9Nq



Supplementary Figure S1. Visualisation of sequence quality over length. Upper four plots are example outputs of FastQC that summarises the per base sequence quality of illumine reads. The bottom two plots are example outputs of NanoPlot that summarises the average read quality against length for HiFi reads.



Supplementary Figure S2. Comparison of metagenomic samples based on Bray-Curtis dissimilarity of 16S rRNA gene composition. a) Dendrogram generated from Bray-Curtis dissimilarity matrix of samples' community composition at a genus level. b) Non-metric multi-dimensional scaling ordination of Bray-Curtis dissimilarity of samples' community composition at a genus level.

Chapter III

Niche partitioning of the ubiquitous and ecologically relevant NS5 marine group

Taylor Priest, Anneke Heins, Jens Harder, Rudolf Amann and Bernhard M. Fuchs

Manuscript published in ISME Journal

Contribution of the candidate in % of the total work

Experimental concept and design – 60%

Experimental work/acquisition of experimental data – 50%

Data analysis and interpretation – 80%

Preparation of figures and tables – 100%

Drafting of the manuscript – 80%

www.nature.com/ismej

ISME

ARTICLE OPEN

Check for updates

Niche partitioning of the ubiquitous and ecologically relevant NS5 marine group

Taylor Priest[1], Anneke Heins[1], Jens Harder[1], Rudolf Amann[1] and Bernhard M. Fuchs[1]✉

Niche concept is a core tenet of ecology that has recently been applied in marine microbial research to describe the partitioning of taxa based either on adaptations to specific conditions across environments or on adaptations to specialised substrates. In this study, we combine spatiotemporal dynamics and predicted substrate utilisation to describe species-level niche partitioning within the NS5 Marine Group. Despite NS5 representing one of the most abundant marine flavobacterial clades from across the world's oceans, our knowledge on their phylogenetic diversity and ecological functions is limited. Using novel and database-derived 16S rRNA gene and ribosomal protein sequences, we delineate the NS5 into 35 distinct species-level clusters, contained within four novel candidate genera. One candidate species, "*Arcticimaribacter forsetii* AHE01FL", includes a novel cultured isolate, for which we provide a complete genome sequence—the first of an NS5—along with morphological insights using transmission electron microscopy. Assessing species' spatial distribution dynamics across the Tara Oceans dataset, we identify depth as a key influencing factor, with 32 species preferring surface waters, as well as distinct patterns in relation to temperature, oxygen and salinity. Each species harbours a unique substrate-degradation potential along with predicted substrates conserved at the genus-level, e.g. alginate in NS5_F. Successional dynamics were observed for three species in a time-series dataset, likely driven by specialised substrate adaptations. We propose that the ecological niche partitioning of NS5 species is mainly based on specific abiotic factors, which define the niche space, and substrate availability that drive the species-specific temporal dynamics.

The ISME Journal (2022) 16:1570–1582; https://doi.org/10.1038/s41396-022-01209-8

INTRODUCTION

An ecological niche is defined as a specific set of conditions (environmental and biotic interactions) that allow a population to perform its evolutionarily adapted function and as a result, persist or grow [1, 2]. Although a long-standing concept in ecology, niche theory has only recently been incorporated in the study of marine microbial populations [3–6]. Such studies have either focused on adaptations to specific conditions across environments [4] or on specific functional adaptations within an environment, e.g. specialised substrate utilisation [7]. However, to obtain a more detailed understanding on microbial populations' niches, an in-depth analysis on the adaptation to conditions across different spatial and temporal scales in combination with an assessment of ecological function is needed.

Microbial populations exhibit distinct distribution patterns across the world's oceans, which are most influenced by depth [8] and changes in temperature [9, 10] and salinity [11]. However, the effect these have on microbial populations varies. Although some appear to be ubiquitously distributed, such as the SAR11 or *Prochlorococcus* Clade, further analysis has shown that distinct genetic variations exist, resulting in ecotypes that are driven by environmentally mediated selection processes [9, 12]. Within specific environments, microbial populations also exhibit distinct dynamics that are driven by temporally derived shifts, such as seasons [6]. This is particularly evident with heterotrophic microbes in the *Bacteroidetes* phylum, that show recurrent and potentially predictable, seasonal dynamics driven by substrate availability [7, 13, 14]. From these studies, it is clear that conditions and resources influence microbial populations, however, to what extent do these determine niches?

In this study, we phylogenetically and ecologically characterise members of the NS5 marine group (referred to as NS5 from hereon) and subsequently identify the key niche-determining factors over spatial and temporal scales. The NS5 was selected as it represents a ubiquitous and abundant group of the *Flavobacteriia* class for which our knowledge on phylogeny and function is limited. Since the name was introduced 14 years ago, from a study describing high local and temporal diversity of *Flavobacteriia* in the North Sea [15], NS5-classified sequences have been recovered from across the world's marine water masses, ranging from semi-enclosed seas in tropical regions [16, 17] to the Antarctic peninsula [18] and North Pacific oxygen minimum zone [19]. They are frequently reported as one of the most abundant groups of *Flavobacteriia* from studies using 16S rRNA gene analysis [17, 20] and fluorescence in situ hybridisation (FISH) cell counts [21] (referred to as VIS1 in that study). The VIS1 clade, which represents only a fraction of the NS5, was reported to reach $29 \pm 3 \times 10^3$ cells ml^{-1} in the Arctic province of the North Atlantic. Members of the NS5 have been shown to associate with spring phytoplankton blooms [13, 22, 23] and increasing chlorophyll *a* concentrations

[1]Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, Bremen, Germany. ✉email: bfuchs@mpi-bremen.de

Received: 28 July 2021 Revised: 24 January 2022 Accepted: 2 February 2022
Published online: 15 February 2022

[21, 24], however they are typically more prominent in early bloom stages or are more tightly coupled to the fluctuations in flagellate abundance [13]. A study conducted in the South Sea of Korea concluded that NS5 was a good indicator species for coastal waters [25] whilst they were also shown to be linked to eutrophication in coastal bays of Vietnam [22]. In contrast, other findings have indicated NS5 sequence abundance maxima to occur in winter [26] and a dominance of NS5 affiliated sequences in open ocean Arctic waters [20, 21], which motivated us to here provide novel data from the Fram Strait region. From these studies, it is clear that the NS5 may represent a diverse group of bacteria with different ecological niches.

With a global perspective, we here characterise the NS5 based on (1) phylogenetic analysis using 16S rRNA gene and ribosomal protein tree reconstructions, (2) functional descriptions using MAGs and a complete genome of a cultured isolate, (3) spatiotemporal distribution patterns of species-level cluster representatives and (4) visual identification of cells in culture using transmission electron microscopy (TEM) and in the environment using catalysed reporter deposition-fluorescence in situ hybridisation (CARD-FISH). As a result, we identify and characterise 35 species assigned to four novel candidate genera, which we have named *Candidatus* Marisimplicoccus (NS5_A), *Candidatus* Marivariicella (NS5_B), *Candidatus* Maricapacicella (NS5_D) and *Candidatus* Arcticimaribacter (NS5_F), for each of which, we propose a candidate type species based on a genome voucher.

MATERIALS AND METHODS

Fram Strait sampling and sequencing

Seawater samples were collected at the deep chlorophyll maximum (DCM) layer from 11 stations across the Fram Strait region in July and August 2018 during the PS114 Polarstern cruise, as described previously [27]. Seawater was fractionated using filtration and DNA extracted using a modified SDS-based extraction method after Zhou et al. [28]. Metagenomes were generated from the 0.2–3 µm fraction using the HiSeq 3000 (Illumina) and Sequel II (PacBio) platforms, as described previously [27] (Supplementary Material 1).

Generation of MAG dataset

The metagenomic reads from the Fram Strait samples were assembled (using Megahit [29] for Illumina reads and MetaFlye [30] for PacBio reads), binned (using Concoct [31], Metabat2 [32], Maxbin2 [33] and DAStool [34]) and manually refined as described previously [27], resulting in MAGs identified in this study by the prefix "FRAM18_". Estimation of genome completion and contamination was determined using CheckM v1.1.2 [35]. MAGs belonging to the NS5 were identified by 16S rRNA gene phylogeny (Supplementary Material 1) and additionally assigned to taxonomic groups in the Genome Taxonomy Database (GTDB) (Release 89) using the classify_wf pipeline of GTDB-tk v1.0.2 [36, 37]. The dataset was expanded by retrieving all species-representative assemblies within the assigned GTDB taxa along with MAGs from two additional 0.2–3 µm marine microbial metagenomic datasets (Bioproject accessions: PRJEB28156 [38, 39], PRJEB43746) that had been assigned the same GTDB taxonomy (Supplementary Tables S1 and S2). The resulting dataset was de-replicated using FastANI v1.9 [40] with a cut-off threshold of 95%.

Helgoland NS5 isolate AHE01FL: sampling, isolation and genome sequencing

Seawater from the long-term ecological research station Helgoland Roads (54°11′03″N, 7°54′00″E) was sampled on the 28th April, 2016 and serially diluted with artificial seawater [41]. An inoculum of 2.6 nl, statistically containing three cells, was grown in an oligotrophic HaHa medium with the addition of vitamins [42] in the dark at 12 °C. After several transfers and another dilution to extinction series, purity controls confirmed that the culture contained a pure strain. It was maintained by transfers every 3 months. Growth in HaHa100V medium [42] yielded a turbid, orange-coloured culture and provided biomass for DNA extraction that was performed according to Zhou et al. [28]. Genome sequencing was performed by the Max Planck-Genome-centre Cologne, Germany (https://mpgc.mpipz.mpg.de/home/) using Sequel I (PacBio) and HiSeq 2500 (Illumina) platforms. Circular long read sequences from PacBio were assembled using Canu v2.1 [43] whilst short Illumina reads were assembled using Spades v3.13.2 (parameters: -isolate) [44]. The contigs from both datasets were aligned and assembled together in Geneious Prime v2019.1.3 (https://www.geneious.com) before a final round of error-correction using the Illumina reads as a reference. The assembled genome was submitted to EMBL-EBI and assigned the name "*Flavobacteriaceae* bacterium AHE01FL" with the taxid 2820661 and in this study, is named "Iso_AHE01FL".

Helgoland NS5 isolate AHE01FL: cell visualisation

To accurately determine cell morphology of Iso_AHE01FL, TEM was used. An aliquot of the liquid culture was retrieved and fixed with 25% glutaraldehyde (EM grade Science Services) for 1 h at room temperature followed by centrifugation (5 min at 21,100 × *g*) and resuspension in the growth media (HaHa 100 V medium). An aliquot of this resuspension was pipetted onto a Formvar coated 400-mesh copper grid and stained with 1% uranylacetate for 5 min before being air dried overnight.

NS5 MAG phylogenetic tree reconstruction

The reconstruction of a phylogenetic tree for species-representative MAGs was performed using a concatenated alignment of 16 ribosomal proteins (L2, L3, L4, L5, L6, L14, L16, L18, L22, L24, S3, S8, S10, S17, S19), following the procedure described by Hug et al. [45]. In brief, Muscle v3.8.15 [46] was used to align amino acid sequences that were subsequently trimmed using TrimAl v1.4.1 [47] and concatenated into a single alignment that was provided as an input to FastTree v2.1.10 [48] (Supplementary Material S1). This workflow was then repeated with the addition of 1275 *Flavobacteriaceae* assemblies from the RefSeq database (Supplementary Material S1). To corroborate the inferred phylogenetic separation of MAGs, average nucleotide identity (ANI) and average amino acid identity (AAI) were calculated using FastANI and CompareM v0.1.1 (https://github.com/dparks1134/CompareM), respectively.

16S rRNA phylogenetic tree construction

16S rRNA gene sequences, longer than 1 kbp in length, were extracted from species-representative MAGs using Barrnap [49]. The sequences were imported to the ARB programme [50], aligned using the SINA aligner [51] and phylogenetically placed into the SILVA 138.1 SSU Ref NR99 reference tree using the parsimony algorithm. The MAG-derived sequences along with 100 of the highest quality NS5 sequences in the SILVA database were used for phylogenetic tree reconstruction. Three tree algorithms were used, RaxML v8.2.8 maximum likelihood (GTR-Gamma rate distribution model, rapid bootstrap algorithm, 100 repetitions) [52], neighbour-joining (Jukes-Cantor's substitution model, 1000 bootstrap repetitions) and Parsimony v3.6, each with two different positional variability conservation filters, a 30% for all *Flavobacteriia* and the "termini" filter provided with the ARB SILVA database. A consensus tree was constructed from these six input trees and groups that remained stable throughout all tree methods were designated.

Probe design and environmental cell visualisation

CARD-FISH [53, 54] probes could be designed in ARB for two genus-level clades, NS5_A and NS5_F (Supplementary Table S2). Optimal hybridisation conditions were determined by testing on filtered pelagic water samples from the Fram Strait region [27], the same samples used to generate the "FRAM18_" MAGs. The probes were subsequently applied to five samples from that dataset to obtain information on morphology and cell count data. More detailed information is provided in Supplementary Material S1.

Global distribution of NS5 subgroups, MAGs and their correlation to physical parameters

The distribution of NS5 members was determined by recruiting metagenomic reads from the Tara Oceans dataset (ENA study accessions: PRJEB1787, PRJEB9740) [55] to species-representative MAGs using BBMap v38.73 [56], with a 99% identity threshold (minid = 99, idfilter = 99). In total, 122 surface water, 95 DCM and 47 mesopelagic metagenomes were used. To provide comparability between samples, the number of mapped reads was converted to reads per kilobase per million (RPKM) [57]. The generated data were imported into RStudio [58] and visualised using the

packages *rnaturalearth* [59], *sf* [60] and *ggplot2* [61]. To check the accuracy and provide support for the RPKM values, comparisons were made to genome coverage of mapped reads, cell counts (see "Probe design and environmental cell visualisation") and another, more robust metric, the truncated average depth (TAD) [62], detailed information is provided in Supplementary Material S1.

To determine the effect of abiotic characteristics on NS5 species distribution, physical parameter measurements (depth, chlorophyll *a*, nitrite, nitrate + nitrite, oxygen, phosphate, salinity and silicate) of Tara Oceans samples were obtained from ENA-EBI. Scatter plots of RPKM values across physical parameters were produced using ggplot2 and Pearson's correlation analysis performed using log transformed parameter values.

Seasonal dynamics of species-representative MAGs

Temporal dynamics of species-representative MAGs was determined by read recruitment of oligotypes from a multiyear time-series dataset [63] sampled at Helgoland Roads, North Sea (Supplementary Material S1). Recruitment was performed by BBMap with a 100% identity threshold (minid = 100, idfilter = 100). The distribution dynamics, based on relative abundance of oligotypes taken from the original manuscript, were visualised using the *vegan* [64] and *ggplot2* packages in RStudio.

Functional characterisation

The presence of major metabolic pathways was determined using KofamKoala [65] and RAST v2.0 [66]. For each MAG, initial gene prediction was performed by Prokka v1.14.6 [67]. Carbohydrate-active enzymes (CAZymes) were predicted using a combination of HMMscan against the dbCAN v9 database [68] (*E*-value threshold: 1E−5) and Diamond blastp v0.9.14 [69] against the CAZy database (release 07312020) [70] (*E*-value threshold: 1E−20, parameters: -more-sensitive –query-cover 40 –id 30 –k 15). Sulfatases were annotated by blastp search against the SulfAtlas v1.3 database [71] (*E*-value threshold: 1E−4) and HMMscan against the Pfam sulfatase family PF00884 (*E*-value threshold: 1E−5). Peptidases were identified by blastp search against the MEROPS database [72] (*E*-value threshold: 1E−4). TonB-dependent transporters (TBDTs) were predicted by HMMscan against TIGRFAM profiles TIGR01352, TIGR01776, TIGR01778, TIGR01779, TIGR01782, TIGR01783, TIGR01785, TIGR01786, TIGR02796, TIGR02797, TIGR02803, TIGR02804, TIGR02805, TIGR04056 and TIGR04057 (*E*-value threshold: 1E−10). SusD genes were identified by HMMscan against the Pfam profiles PF12741, PF12771, PF14322, PF07980. Annotations of carbohydrate esterases, carbohydrate binding modules, glycoside hydrolases (GH) and polysaccharide lyases (PL) were designated correct only if both the dbCAN and CAZy annotations agreed. Annotations were combined into a single "gene_table" for each MAG. To identify potential polysaccharide utilisation loci (PULs), text searches were performed in the "gene_table" for regions on contigs that contained either a SusC/SusD gene pair with two or more degradative CAZymes or contained at least three substrate utilisation genes in close proximity (maximum of 6 genes in between each). PULs were manually inspected and visualised using the *gggenes* [73] and *ggplot2* packages in RStudio.

The composition of CAZyme, sulfatase, peptidase and TBDT gene families for all MAGs was subsequently converted to a Bray-Curtis dissimilarity matrix and used as an input for hierarchical clustering and a non-metric multi-dimensional scaling analysis, using the hclust and metaMDS functions of the *vegan* package in RStudio. The visualisation of the analyses was carried out using the *ggplot2* and *ggdendro* [74] packages.

SusC/SusD protein trees

Amino acid sequences of SusC/SusD genes identified in PULs were extracted and used for tree calculation. Additional SusC/SusD sequences were included from previously published marine flavobacteria MAGs [38] and cultured isolates [75]. Multiple sequence alignments were calculated using MAFFT v7.310 [76] with L-INS-I and trees calculated using FastTree. Trees were visualised and annotated in iTOL v4 [77].

RESULTS

Seven species-representative MAGs retrieved from Fram Strait metagenomes were identified as members of the NS5 marine group through 16S rRNA gene analysis and assigned to four different genera within the GTDB database (MED-G11, GCA-002723295, MS024-2A, UBA7428). The GTDB species-representatives within these groups along with MAGs from two other metagenome datasets were acquired (Supplementary Tables S1 and S2). In addition, we sequenced and assembled a complete genome of an isolate retrieved from surface seawater at Helgoland Roads, North Sea in 2016. The derived dataset of assembled genomes was de-replicated at a 95% ANI threshold, resulting in 35 species-level clusters that provided the foundation for a detailed phylogenetic and ecological characterisation of the NS5 marine group (Supplementary Table S1).

Phylogenetic analysis of the NS5 marine group

Phylogenetic tree reconstruction using 16S rRNA genes from NS5 species-representatives and sequences from the SILVA 138 database resulted in six distinct clusters being formed (NS5_A–NS5_F) (Fig. 1a). However, the MAG sequences were positioned only within five of the ribosomal protein-based clusters (not NS5_E), indicating that part of NS5's diversity is not yet captured by MAGs. Minimum intra-group 16S rRNA gene sequence similarity varied from 93% in NS5_D to 97.0% in NS5_F whilst the median values ranged from 94.5% in NS5_B to 98.9% in NS5_F. The lower values observed were typically a result of only a few sequences, with the majority of minimum values being >94.5% and median values >96.4% and therefore, in agreement with genus-level thresholds [78].

Reconstruction of a MAG-based ribosomal protein tree (Fig. 1b) revealed five distinct groups that corresponded to clusters in the 16S rRNA gene tree. The number of species-representative MAGs in each cluster ranged from 17 in NS5_D to 1 in NS5_C. Genomic comparisons between MAGs revealed intra-cluster average AAI values of >65% and inter-cluster values of <65%, further supporting the delineation of groups at the genus-level [79]. The coherence and stability of the clusters was additionally confirmed by phylogenetic tree reconstruction at the family level (Supplementary Fig. S1). Due to the genetic coherence, the defined clusters will now be referred to as genera. The cultured isolate, Iso_AHE01FL, belonged to the NS5_F genus. In order to provide an indication on genetic conservation, the species-representative genomes from NS5_F were aligned to the complete isolate genome and a visualisation provided on the conserved syntenic gene blocks identified (Supplementary Fig. S2).

Clear distinctions between genera were evident with respect to genome size and GC content (Supplementary Fig. S3). The average genome size of the three most complete MAGs from each genus were 2.05 Mbp for NS5_F, 2.02 Mbp for NS5_D, 1.82 Mbp for NS5_B and 1.17 Mbp for NS5_A, whilst the GC content of NS5_A and _B representatives was ~30% compared to 36 – 37% in NS5_D and _F.

Cell visualisation

Following the design and optimisation of CARD-FISH probes (Supplementary Material S1) for the NS5_A and NS5_F genera (Supplementary Table S3), cells were visualised on filtered seawater samples from the Fram Strait region [27] (Fig. 2). Probe design for the other genera was unsuccessful due to sequence similarities with neighbouring taxa. Hybridised cells visualised using the NS5_A probe were of a small coccoid shape with a diameter of ~0.5 µm. Those identified with the NS5_F probe were rod-shaped cells with a length of 0.5–1.5 µm and width of ~0.5 µm. Enumeration of FISH signals that overlapped with a nucleic acid stain (DAPI) revealed similar peak counts for both NS5_A, 1.70×10^4 cells ml^{-1}, and NS5_F, 1.76×10^4 cells ml^{-1} (Supplementary Fig. S4).

TEM on the Iso_AHE01FL revealed rod-shaped cells with a length of 0.5–1 µm and width of <0.5 µm (Fig. 2), in agreement with the observations on environmental samples, based on FISH.

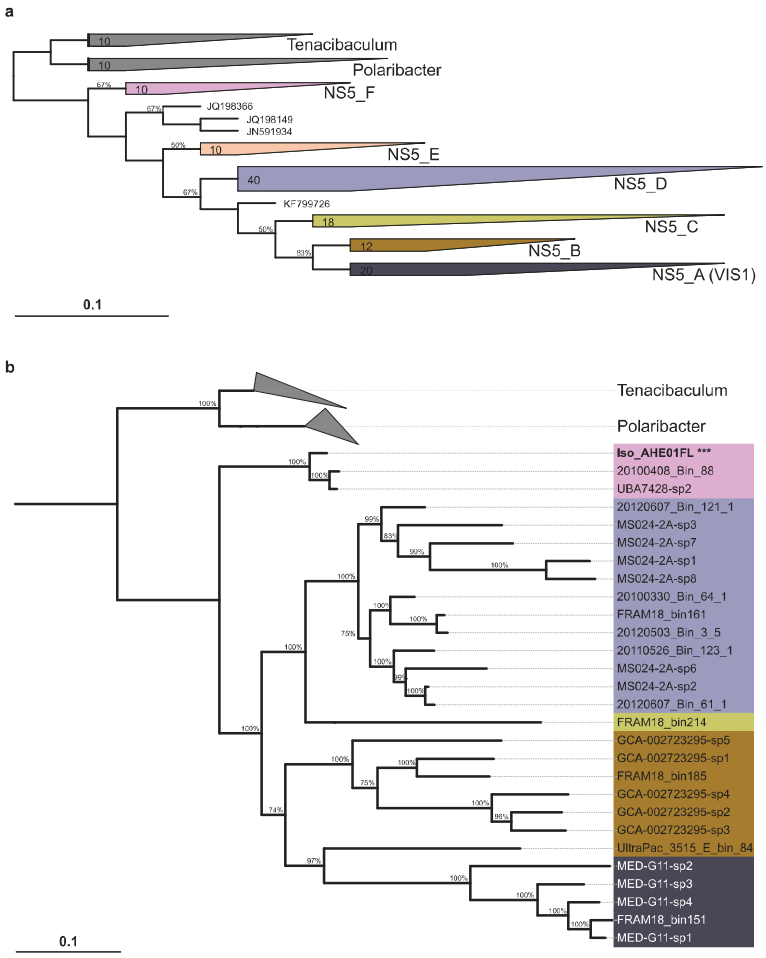


Fig. 1 Phylogenetic tree reconstruction of the NS5 Marine Group. a 16S rRNA gene tree constructed using MAG sequences, from this and previous studies and the GTDB database, and 100 sequences classified as NS5 Marine Group in the SILVA 138 database. The tree represents a consensus from six input trees, constructed using three different algorithms, RaxML, Neighbour-joining and parsimony, with two different positional variability filters. **b** Ribosomal protein tree generated from a concatenated alignment of 16 proteins identified within NS5 species-representatives MAGs and genomes of Polaribacter and Tenacibaculum retrieved from the NCBI RefSeq database. The cultured isolate, Iso_AHE01FL, is highlighted in bold and with ***.

Global distribution of NS5 genera, MAGs and their correlation to physical parameters

The distribution of NS5 genera was determined by read recruitment from Tara Oceans metagenomes to each individual species and subsequently summing the RPKM values (Supplementary Fig. S5). To provide additional support, RPKM values were compared to genome coverage of mapped reads, CARD-FISH cell counts and another, more robust sequence-based metric, the TAD [62] (Supplementary Material S1 and Supplementary Figs. S4, S6 and S7). Based on this, a cut-off threshold of 0.25 RPKM was applied for inclusion in further analysis, which ensured a coverage of >40%. The four genera each exhibited a ubiquitous presence across all oceanic regions in the surface and DCM layers, although the NS5_F genus was less widespread in the DCM than surface. All genera showed lower RPKM values in the mesopelagic than DCM layer. The magnitude of RPKM values observed for NS5_B was six-fold lower than for the other genera. Variations in distribution patterns were evident, with the NS5_D and NS5_F reaching higher RPKM values in Arctic and geographically connected areas whilst NS5_A appeared more prevalent in specific locations, such as the North Atlantic and Chilean upwelling system. These patterns were further confirmed by grouping samples into oceanic regions (Supplementary Fig. S8).

On a species-level, an almost universal distribution pattern with depth was identified, with all but two species exhibiting highest RPKM values in surface waters (<30 m) (Supplementary Fig. S9). The two contrasting species were MS024-2A_sp7 (NS5_D), which peaked between 100–200 m, and MED-G11_sp2 (NS5_A), which peaked at ~300 m depth. Additionally, several species exhibited a bimodal peak, with highest values in surface waters but an

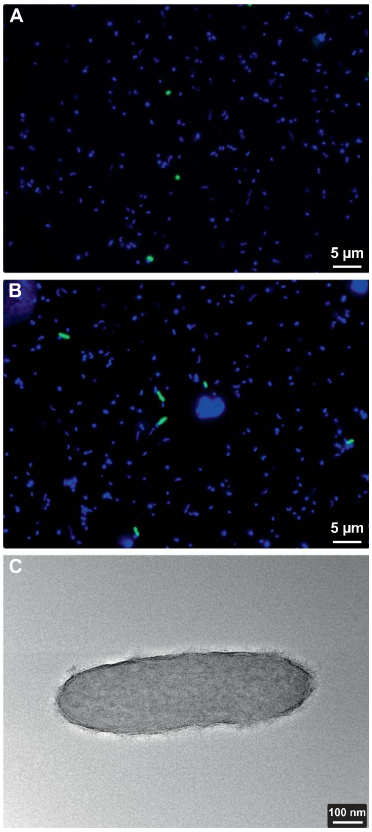


Fig. 2 Visualisation of cells from NS5_A and NS5_F. Environmental cells hybridised using CARD-FISH probes targeting the NS5_A (**A**) and NS5_F (**B**). FISH probe signals are shown in green and DNA stain in blue. **C** Transmission electron microscopy image of the isolate, Iso_AHE01FL, in the NS5_F.

additional, smaller peak in RPKM observed in mesopelagic depths, such as FRAM18_bin185 (NS5_B). Asides from depth, the geographical distribution patterns of species within and between genera varied (Supplementary Figs. S10–S14) which typically reflected distinct dynamics in relation to temperature (Supplementary Fig. S15), salinity (Supplementary Fig. S16) and oxygen (Supplementary Fig. S17). However, no clear patterns were evident with respect to nitrate + nitrite, phosphate, silicate or chlorophyll *a*. The geographical distribution patterns of species could be categorised into three types.

The first, encompasses species-representatives with higher RPKM values in a specific geographical region, e.g. the Mediterranean and Red Sea for MS024-2A_sp5 (Fig. 3). As a result, representatives of this type exhibited narrow peaks in RPKM values in relation to abiotic conditions, e.g. for MS024-2A_sp5 at ~15 °C and ~38 psu. The second pattern is represented by changes in RPKM values with latitude, e.g. higher values in the Arctic for species UBA7428_sp2 (Fig. 3) and all species in NS5_F or in temperate regions for GCA-002723295_sp2 in NS5_B (Supplementary Fig. S11) and MS024-2A_sp7 in NS5_D (Supplementary Fig. S13). These species typically exhibited peak RPKM values within a defined range of each abiotic condition. For example, the Arctic-preference distribution of NS5_F species was related to peaks in RPKM values at temperatures <5 °C, oxygen concentrations >300 µM and salinities <33 psu whereas the temperate-preference distribution of UltraPac_E_bin_84_1 was related to peaks across a wide range of temperatures, 12–30 °C, and at salinity values of 33–38 psu. Lastly, the remaining species exhibited an unclear distribution pattern, either due to below-threshold RPKM values in most samples or comparable RPKM values in samples without a clear pattern, e.g. GCA-002723295_sp1 in NS5_B (Fig. 3). Representatives of this last distribution type, as could be expected, showed a lack of or an inconsistent pattern with shifts in abiotic conditions. Species RPKM values across all Tara Oceans samples are provided in Supplementary Table S4.

Seasonal dynamics of NS5 species

By performing read recruitment analysis of 16S rRNA gene oligotypes from a previously published time-series dataset, we were able to visualise the temporal dynamics of six NS5 species-representatives at Helgoland Roads, German Bight. Each of the identified oligotypes exhibited distinct and recurrent temporal dynamics over three consecutive years (Fig. 4), with three also showing a successional pattern from spring to summer. This succession began with 20100330_Bin_64_1 (NS5_D) that peaked from early to late spring (up to 4.5% of the community), followed by FRAM18_bin181 (NS5_F) in late spring and FRAM18_bin161 (NS5_D) that also peaked in late spring but persisted throughout summer (up to 3.5% of the community). Although the isolate, Iso_AHE01FL, was recovered from Helgoland Roads, the respective oligotype was present in low relative read abundance throughout the annual cycle (<0.1% of the community).

Functional characterisation

In order to assess functional differences across the NS5 genera, the gene annotations that were consistent across the three most complete MAGs in each genus were compared (referred to as genus-level values from hereon). The necessary genes for glycolysis, gluconeogenesis, the pentose phosphate pathway, the tricarboxylic acid cycle and for the major components of the electron transport chain were identified in all genera, confirming an aerobic heterotrophic metabolism. An additional unifying feature was the presence of a green-light proteorhodopsin (PR), which is not found in low light conditions. Mechanisms for nitrogen and phosphorous metabolism were conserved across all groups and restricted to an ammonium transporter (Amt family) and nitrogen response regulatory proteins (e.g. NtrC) along with the ability to build and hydrolyse long chain polyphosphates with a polyphosphate kinase (*ppk*) and an exopolyphosphatase. In addition, all species contained a glycogen synthase gene, indicating the capacity to use glycogen as a storage molecule. In contrast, genes related to sulfur metabolism were not conserved across genera, with only the NS5_B harbouring the capacity for assimilatory sulfate reduction. The ability to synthesise riboflavin was conserved whilst all genera lacked the genes required for biotin, thiamine and vitamin B12 synthesis. In contrast, significant differences were evident with respect to substrate acquisition and degradation potential between NS5 genera.

The annotation of CAZymes varied considerably across genera and species (Supplementary Tables S5 and S6). The number of glycoside hydrolase (GH) genes across all species-representatives ranged from 0–12 per Mbp (Fig. 5 and Supplementary Table S5), whilst genus-level values ranged from 9 per Mbp in NS5_D and _B to 5 per Mbp in NS5_A. There were also clear distinctions in the composition of conserved and non-conserved GH gene families within each genus (Supplementary Fig. S18). The NS5_D harboured eight conserved gene family annotations compared to four in NS5_F, three in NS5_B and one in NS5_A. There were no

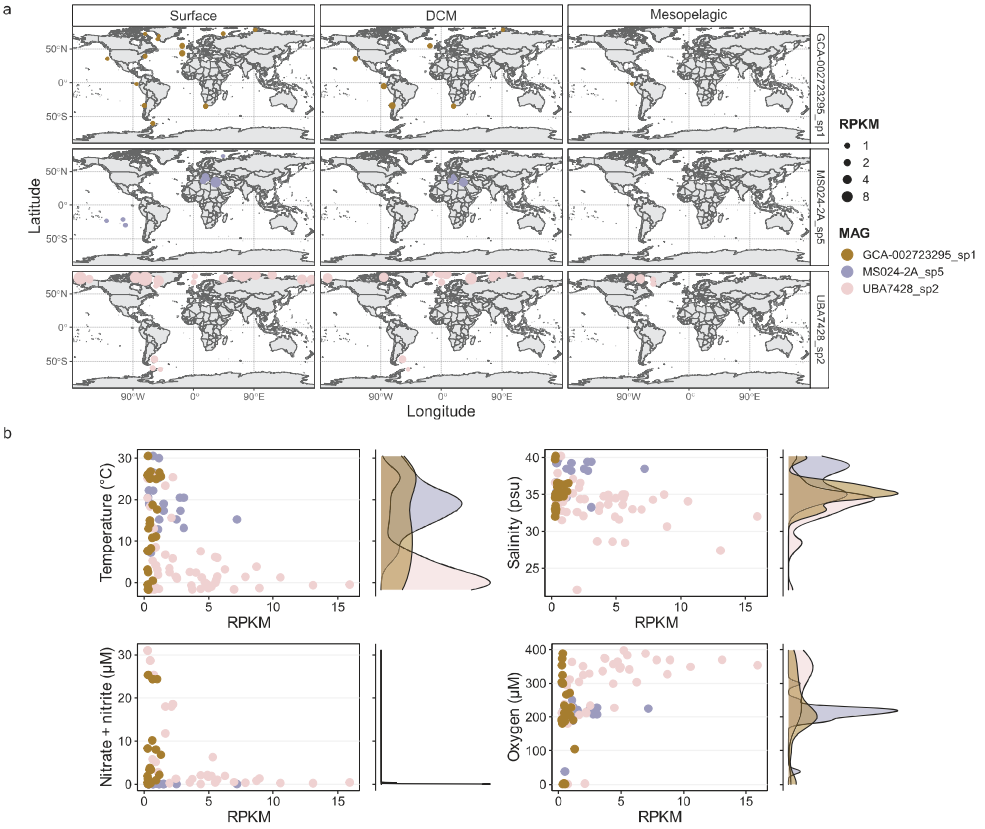


Fig. 3 Three select species that represent different distribution types observed across the NS5 and their dynamics in relation to abiotic conditions. RPKM values were calculated based on read recruitment from Tara Oceans metagenomes to species-representative MAGs using BBMap with a 99% identity threshold. A minimum threshold of 0.25 RPKM was applied which ensured a minimum genome coverage of 40%. **a** Global distribution and **b** dynamics in species RPKM values across abiotic conditions. Within the scatter plots, each point represents a Tara Oceans sample where the species RPKM value was >0.25. Alongside each scatterplot is a density diagram showing the distribution of points.

universally conserved gene families. However, two conserved gene families were shared between the NS5_B, _D and _F, including a GH16_3 (β-1,3-glucanase) and a GH3. Conserved gene families specific to a single genus included GH29 and GH95 (both known as α-fucosidases) in NS5_D and GH113 (β-mannanase) in NS5_F. The large range in non-conserved GH gene family annotations across genera indicated a varying degree of substrate-metabolic diversity on the species-level (Supplementary Fig. S18). Most notable was the diversity within NS5_B, with 26 different GH gene families or sub-families. This also provided evidence that potential substrates, not conserved at the genus-level, are shared between species of different genera. For example, annotations for α-fucosidases were not restricted to NS5_D, but also found in some species of NS5_B (GH151) and NS5_F (GH107). The presence of GH16_3, GH18 and GH20 genes across species from all genera indicated a shared potential to degrade β-1,3-glucans, such as laminarin, and β-hexosamines, such as peptidoglycan. In addition, a number of annotations were unique to some species within a single genus, including GH43_1 (β-xylosidase/α-L-arabinofuranosidase) and GH142 (β-L-arabinofuranosidase) in NS5_B, GH13_31 (α-glucosidase), GH28 (α-L-arabinofuranosidase) and GH28 (poly-/rhamno-galacturonase) in NS5_D and GH144 (β-1,2-glucosidase) in NS5_F. Further comparisons on GH gene family annotations revealed that each species' composition is unique (Supplementary Table S6).

A major process in carbohydrate catabolism in heterotrophic microbes involves glycan transport into the cell, a process mediated by, among others, TBDTs. The number of annotated TBDTs at the genus-level ranged from 7–9 per Mbp (Table 1) and at a species-level, from 4–13 per Mbp (Fig. 5 and Supplementary Table S5). In addition to TBDTs, the composition of all transporters was compared between genera (Supplementary Fig. S19). There were 23 universal transporters, including for vitamins and metals (Vitamin B12, zinc and magnesium), peptides (Di-tripeptide and D-serine) and carbohydrates (sodium/glucose, L-idonate, high-affinity gluconate and sugar SemiSWEET). In addition, 11 transporters were shared between more than one genus whilst 23 were unique to a single genus. The NS5_B and NS5_D both shared transporters indicative of more versatile metabolisms, including fatty acid and C4-dicarboxylate TRAP transporters.

Distinct differences were also observed for sulfatase and peptidase gene annotations. The number of sulfatases ranged considerably, from 2 to 24 per Mbp across species (Fig. 5) and on the genus-level, from 4 per Mbp in NS5_F to 13 per Mbp in NS5_B

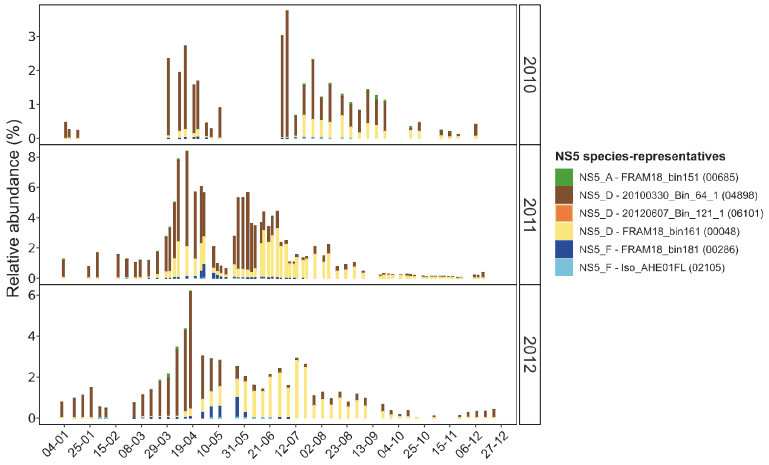


Fig. 4 Temporal dynamics of NS5 species-representatives in surface waters at Helgoland Roads, North Sea. Distributions were obtained by recruiting oligotype representatives from a previously published dataset (62) to each species-representative MAG using BBMap with a 100% identity threshold. Next to each species-representative name, the original oligotype number is provided for direct comparison with previous dataset. Only recruitments successful with a 100% identity threshold are included. The relative abundances for each oligotype were taken from the original publication.

(Table 1). In comparison, peptidases were more consistent at the genus-level, ranging from 7 to 8 per Mbp (Table 1) and exhibited a narrow range across species, 6–11 per Mbp (Fig. 5).

To provide an additional perspective on substrate preferences, the ratio of GH genes to other substrate utilisation genes was calculated (Table 1), a metric that has previously been employed for other flavobacterial groups [7]. The NS5_A consistently had the lowest ratios of GH genes and was the only genus to harbour less GH genes than peptidases, 1:1.6. The ratio of CAZymes:sulfatases also varied across genera, with NS5_F being the only genera to contain more GH genes than sulfatases (Table 1).

Comparing the gene repertoire of CAZymes and peptidases across all species-representatives through a dissimilarity distance matrix approach, resulted in a clustering of species based on phylogeny (Fig. 6). This suggests that the substrate utilisation potential is primarily determined through evolution, and not an adaptation to habitats based on lateral gene transfer. However, refining the dataset to only contain specific sets of genes, e.g. only CAZymes, resulted in less coherent phylogeny-based clustering, although the effects across genera varied depending upon the chosen gene set (Supplementary Fig. S20).

Polysaccharide utilisation loci (PULs) and SusC/SusD protein trees

PULs are genetic clusters of functionally related genes that are involved in the binding and cleavage of polysaccharides and subsequent uptake of oligosaccharides into the cell [80]. Canonical PULs are those containing degradative CAZymes and a SusC/SusD gene pair [80], which provides the transport mechanism for large oligosaccharides across the outer membrane. However, atypical or non-canonical PULs, lacking the SusC/SusD gene pair but still consisting of numerous degradative CAZymes, have also been described [81]. PULs are typically specific towards certain polysaccharides and can thus provide valuable information on variations in substrate metabolism across species, even if the SusC/SusD gene pair is absent. The total number of PULs across NS5 species ranged from 0 to 6, with an almost complete absence in NS5_A (Table 1 and Supplementary Table S7).

PUL structures were highly diverse across species, but four examples of conserved gene colocalisations were identified at the genus-level, two in NS5_D and two in NS5_F (Fig. 7). In NS5_D, one conserved colocalisation consisted of several GH29 genes (α-fucosidases), a single GH33 (sialidase) and/or GH3 gene and at least two sulfatases. Additionally, species within NS5_D harboured a PUL containing a solute-binding protein, C4-dicarboxylate TRAP transporter, 2,3-diketo-L-gulonate TRAP transporter, uronate isomerase and mannonate dehydratase which was accompanied by a GH95 and fructokinase gene in many of the representatives. Such a structure was also identified in one MAG from NS5_B, UltraPac_3515_G_bin_18, and indicates an ability to uptake sugar acids and C4 carbon compounds. In the NS5_F, one conserved colocalisation consisted of several GH16_3 (β-1,3-glucosidase) genes with a GH3 gene, with all but one MAG also encoding for a GH109 (N-acetylhexosaminidase) and galactokinase gene in the same region. In Iso_AHE01FL, this was further supplemented by a sodium/glucose cotransporter, a GH65 and β-phosphoglucomutase gene, providing additional machinery for β-glucan degradation. The second conserved colocalisation in NS5_F consisted of a double SusC/SusD gene pair along with at least two polysaccharide lyase (PL) genes from the PL6, PL7 and PL17 families, suggesting alginate as a potential substrate target (Fig. 7).

The reconstruction of protein trees using SusC and SusD genes additionally confirmed the predicted substrate targets of PULs (Supplementary Fig. S21). For example, the SusC and SusD genes derived from potential alginate-targeting PULs identified in NS5_F species, clustered with those from alginate-targeting PULs of previously recovered MAGs and cultured isolates of *Flavobacteriia*.

Novel candidate genera

Based on the phylogenetic and ecological partitioning of NS5 species, sufficient information has been collected to formally describe four candidate species and genera within the *Flavobacteriaceae* family, *Candidatus* Marisimplicoccus (NS5_A), *Candidatus* Marivariicella (NS5_B), *Candidatus* Maricapacicella (NS5_D) and *Candidatus* Arcticimaribacter (NS5_F). The etymology and

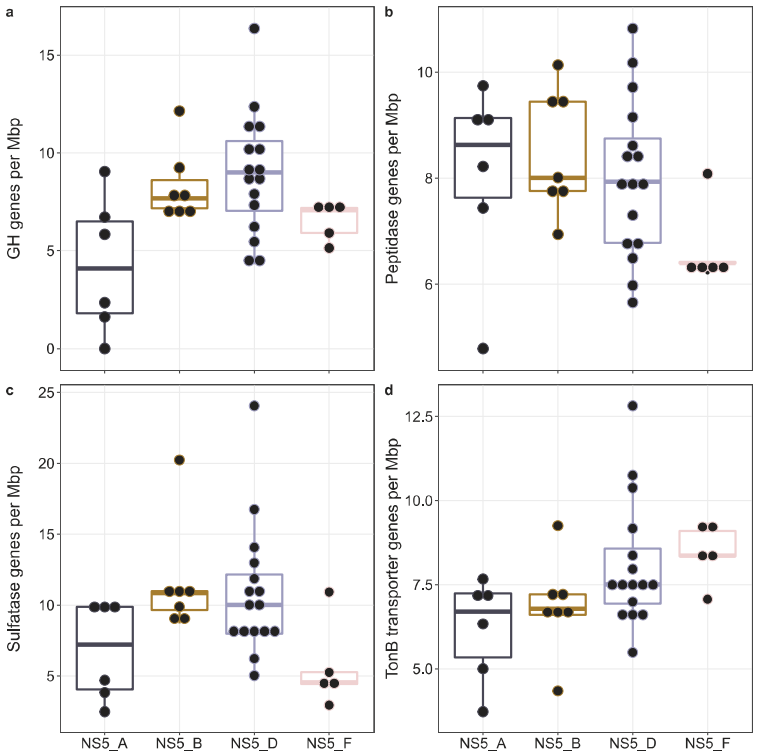


Fig. 5 Summary of substrate utilisation genes annotated in species-representative MAGs. a Number of glycoside hydrolase genes based on agreeing annotations from HMMscan against dbCAN database and Diamond blastp search against the CAZy database. **b** Number of peptidases annotated using blastp search against the MEROPS database. **c** Number of sulfatase genes based on HMMscan against the Pfam sulfatase profile and blastp search against the SulfAtlas database. **d** Number of TonB-dependent transporters based on HMMscan against TIGRFAM TonB profiles.

metabolic descriptions of these are provided in Supplementary Material S1 and Supplementary Figs. S22–S25.

DISCUSSION

The NS5 marine group represent one of the most prevalent groups of marine flavobacteria across the world's oceans yet our knowledge on their phylogenetic diversity and ecological functions is limited. Here, we phylogenetically and ecologically characterise four novel candidate genera within the NS5, each with a candidate representative species. The genera encapsulate 35 distinct species that are distinguishable by genomic characteristics, spatiotemporal distribution patterns and predicted functional potentials, from which we can hypothesise ecological niche partitioning. Furthermore, we present the first complete genome sequence and morphological description of an NS5 isolate, "*Arcticimaribacter forsetii* AHE01FL".

Phylogeny and genomic characteristics

Phylogenetic tree reconstructions revealed four distinct, coherent genera in the NS5, each with multiple species-representatives, which formed a novel branch within the *Flavobacteriaceae* family. It is clear from tree topologies that two additional genera likely also exist, but a lack of representative genomes hinders further investigation.

NS5 species-representatives across genera are distinguishable by their genomic characteristics. The larger, average genome size of NS5_F and NS5_D (2.05 ± 0.19 and 2.02 ± 0.38 Mbp, respectively) are above the median size reported for a dataset of >1200 marine *Bacteroidetes* MAGs with comparable completeness values (1.96 Mbp) [38] whilst NS5_A is within the smallest 10% of genome sizes from that dataset. In general, the genome size of NS5 representatives are smaller than those of related cultured marine *Flavobacteriia*, such as *Polaribacter* (3.1–4.0 Mbp), *Tenacibaculum* (3.2–5.5 Mbp) and *Formosa* sp. *B* (2.7 Mbp). Genome sizes of cultured isolates and MAGs has previously been shown to vary by up to 1 Mbp in *Polaribacter* [7], however, the genome sizes of NS5_F MAGs were comparable to "*Arcticimaribacter forsetii* AHE01FL" (2.03 Mbp) in the same genus. Furthermore, within NS5_D, one of the species-representatives, MS024-2A_sp8, is a high quality draft single cell genome [16] which also has a comparable genome size to MAGs within the same genus. Therefore, the smaller genome sizes are likely not a methodological artefact but reflect differences in the life strategy and ecological role of NS5 compared to the other well described *Flavobacteriia*.

Life strategy and metabolism

The major metabolic pathways and cellular functions were largely conserved across all four newly described candidate genera and

Table 1. Genomic statistics and carbohydrate and peptide-degradation gene repertoire of NS5 genera.

	Genome completeness (%)	Genome contamination (%)	Genome size (Mbp)	GC content (%)	GHs/ Mbp	CAZymes/ Mbp	Peptidases/ Mbp	Sulfatases/ Mbp	TBDTs/ Mbp	SusCs/ Mbp	SusDs/ Mbp	PULs	GH: peptidase	GH: sulfatase	GH: TBDT
NS5_A	81.4	0.3	1.17	30	5	7	8	8	7	2	3	0	1:1.6	1:1.6	1:1.4
NS5_B	97.3	0.5	1.82	30	9	12	8	13	8	2	2	2	1:0.9	1:1.4	1:0.9
NS5_D	99.1	1.4	2.02	37	9	12	7	9	9	4	4	3	1:0.8	1:1	1:1
NS5_F	97.0	0.1	2.05	36	7	10	7	4	9	2	3	2	1:1	1:0.6	1:1.3

Values are derived from the average of the three most complete metagenome-assembled genomes in each genus. Completeness and contamination were estimated using CheckM v1.1.2. Gene groups are shown as per Mbp values.
TBDT TonB-dependent transporters.

indicative of an aerobic photoheterotrophic lifestyle with supplemental energy acquisition through a proteorhodopsin (PR). PRs are light-driven proton pumps that can generate ATP through the proton motive force [82]. PR-mediated photoheterotrophy is widely distributed among marine Archaea and Bacteria inhabiting the photic zone [83], and it has been shown that PR-containing marine flavobacteria have smaller genomes than PR-lacking flavobacteria [84]. Such findings are in-line with the genome sizes reported here. Another key finding of this study is that NS5 species exhibit free-living lifestyles, evidenced by a lack of flagella machineries and gliding motility and a distinct separation of visualised cells from particles. This is in agreement with previous studies that reported an enrichment of NS5 marine group in the free-living fraction (<3 µm) in the North Sea [85] and Southern Ocean [18].

For free-living aerobic heterotrophs, the main source of carbon and nutrients for growth is dissolved organic matter (DOM). As is known for other groups of marine *Flavobacteriia* [84, 86], NS5 species are shown to encode a suite of degradative CAZymes, indicating a specialised capacity for high molecular-weight DOM degradation. The number of GH genes in NS5_B, _D and _F representatives, 7–9 per Mbp, is similar to values reported for MAGs classified in the *Polaribacter* 1-b (9 per Mbp) and 2-a clusters (9 per Mbp) as well as some cultured isolates such as Formosa B (7.7 per Mbp) [86] and *Gramella forsetii* KT0803^T (10.5 per Mbp) [87]. These organisms are known as specialist degraders of algal-derived carbohydrates and are key members of microbial communities following spring phytoplankton blooms [86, 88]. In contrast, the number of peptidases present in NS5 species is considerably lower, 7–8 per Mbp, than in *Gramella forsetii* KT0803^T, 30.5 per Mbp, and *Formosa B*, 25 per Mbp, indicating a reduced capacity for protein hydrolysis. Furthermore, the number of canonical PULs in NS5 species is lower than the average recently reported for a large dataset of marine *Bacteroidetes* MAGs [38]. Such features may be evidence of narrow substrate niches for NS5 species, which has also been suggested previously [16].

Unique substrate-degradation potentials

NS5 species harbour distinct substrate utilisation capacities, with evidence of genus-wide conserved substrate targets also evident in NS5_D and _F. In NS5_F, these consist of laminarin and alginate. Laminarin is a major storage polysaccharide in marine diatoms and a common substrate of marine flavobacteria, based on the widespread presence of PULs [38] and rapid hydrolysis rates in incubations [89]. The NS5_F laminarin-targeting PULs resemble those previously described for *Gramella forsetii* KT0803^T [75], *Gramella* sp. MAR_2010_147 and *Gillisia* spp. Hel1_29 and Hel1_33_132 [75], shown to be upregulated in the presence of laminarin [88]. Alginate, also a widely available polysaccharide, constitutes a key component in brown algal cell walls, representing up to 45% of their dry weight biomass. Alginate-targeting PULs have been identified in a number of marine *Flavobacteriia* [75], however NS5 representatives dominated the alginate PUL cluster in the Helgoland Roads time-series dataset [38]. These PULs contain the Aly (PL6 and PL7) and Oal families (PL15 and PL17) that together, provide the capacity for complete alginate degradation [90, 91]. Additional substrate targets, shared by a minority of, or unique to a single species, also included algal-derived polysaccharides, such as α-fucan (GH107) for "*Arcticimaribacter forsetii* AHE01FL".

A conserved PUL structure identified in species of NS5_D, containing α-fucosidases and sulfatases, suggests a potential to utilise fucose-containing sulphated polysaccharides (FCSP). However, previous FCSP-targeting PULs identified in fourteen isolates of marine *Flavobacteriia*, encoded a more extensive CAZyme gene repertoire, reflecting the complexity of FCSP [92, 93]. It is thus unlikely that NS5_D species utilise FCSPs but instead, cleave fucose groups bound to other carbohydrate structures or

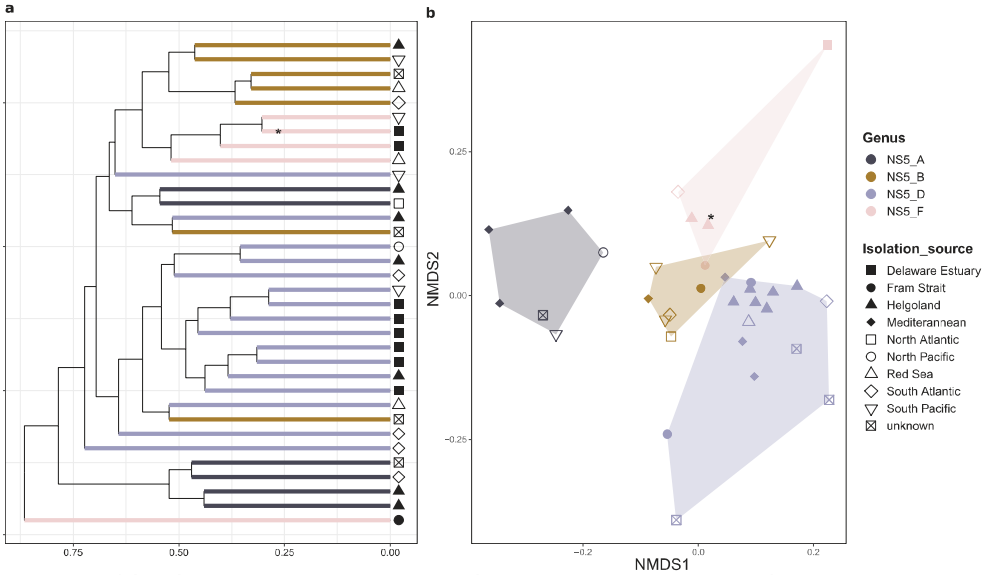


Fig. 6 Comparison of the substrate utilisation gene composition between species-representatives. A dissimilarity matrix was generated from the CAZyme, peptidase, sulfatase and TonB-dependent transporter gene compositions and subsequently used for **a** hierarchical clustering analysis and **b** non-metric multi-dimensional scaling ordination. *indicates the isolate, Iso_AHE01FL.

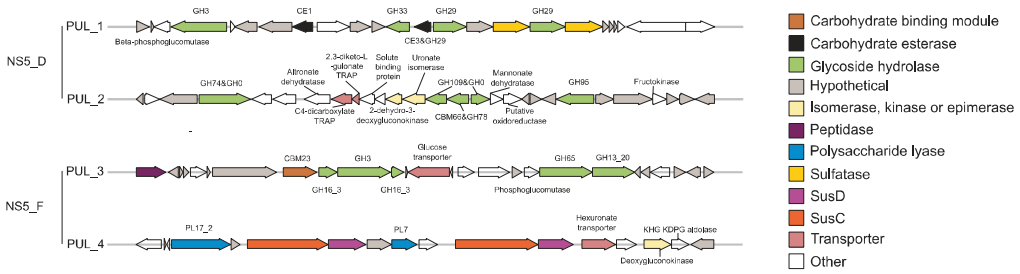


Fig. 7 Polysaccharide utilisation loci containing gene colocalisations that are conserved within candidate genera. The structures presented include two each from the candidate species of NS5_D, 20100330_Bin_64_1, and NS5_F, Iso_AHE01FL. Although the exact PUL structures vary across species, the colocalisation of CAZymes are conserved within all species of the genus. The predicted substrate targets for NS5_D PULs are, PUL_1 = α-fucan and PUL_2 = unclear, and for NS5_F, PUL_3 = laminarin and PUL_4 = alginate.

hydrolyse less complex fucose-containing oligosaccharides in a scavenging-like mechanism. Additionally conserved within NS5_D species, is a genetic loci containing numerous transporters for amino acids, D-xylose, acid sugars and C4 compounds without any degradative CAZymes. Such a structure may be evidence of genome rearrangement to increase the efficiency of gene regulation for substrate acquisition under certain conditions.

A lack of intra-genus conserved gene colocalisations were found in the NS5_B, however, all species contained a PUL targeting bacterial-derived polysaccharides, such as glycogen. In addition, unique PUL structures targeting algal-derived polysaccharides were identified in some species, such as an α-mannose-targeting PUL in GCA-002723295_sp2 (GH95). In NS5_A species, PULs were either absent or low in number and the comparable number of CAZyme to peptidase genes, which was unique to this genus, suggests that proteins are a more important substrate for growth.

Ecological niche partitioning of species

Niche concept has been applied in marine microbial ecology to describe the partitioning of taxa either based on adaptations to specific conditions across environments [4, 6, 94] or adaptations to specialised substrates within an environment [5, 14]. Using time-series data from a coastal ecosystem, it has been shown that populations of *Bacteroidetes* occupy distinct substrate specific niches that drive recurrent temporal dynamics [7, 13, 14]. For the NS5 representatives identified in that dataset, several specific substrate targets were reported, including β-glucan, α-glucan, α-mannan and alginate [38]. We show in this study that these substrates are indicative of different genera. Furthermore, by

using an oligotype dataset from the same time-series, we identified successional-like dynamics for some NS5 species. Those dynamics were also likely driven by substrate utilisation capacity, with the early spring responder, 20100330_Bin_64_1 (NS5_D), encoding for twice as many GH genes and sulfatases as well as a broader diversity of predicted substrate targets than the late spring responder, FRAM18_bin161 (NS5_D). This suggests that substrate may be a major factor in the niche partitioning of these species in that environment. It is important to note however, that other factors, not assessed in this study, likely also contribute to these temporal dynamics, such as grazing by microeukaryotes and viral-induced mortality.

Although substrate may act as a key niche-determining factor for NS5 species in a given environment, we show that species' spatial distribution dynamics across environments and throughout the water column are influenced by distinct shifts in abiotic conditions. Depth, and the associated changes in light and temperature, is well evidenced to structure the vertical distribution of microbial taxa [8]. Such a pattern is also clear for NS5, with nearly all species showing a preference for the upper euphotic zone (<100 m). Adaptations to this environment are evident within the genomes and predicted metabolisms of NS5 species, e.g. the presence of PR and utilisation of HMW-DOM as a substrate that is primarily produced by, or a result of lysis of phytoplankton in the euphotic zone. On geographical spatial scales, studies on microbial biogeography have reported that temperature and oxygen are the strongest correlates to changes in taxonomic and functional composition [95, 96]. The distribution dynamics of NS5 species are in agreement with this, although distinct patterns could also be identified in relation to salinity. It is clear that the niche-determining conditions vary considerably across species, with adaptations to narrow and broad ranges of conditions observed.

The widespread presence of many NS5 species but with distinct preferences for specific environmental conditions are in support of previous theories such as "everything is everywhere but the environment selects" [97] and the microbial seed bank hypothesis [98, 99]. The capacity to survive "everywhere" likely reflects an evolutionary adaptation that resulted in small genomes and cell sizes with advantageous features such as a PR and a potential to utilise widely available substrates. However, it is clear that each species has adapted to a specific set of conditions under which it can proliferate within its defined ecological niche. The factors that determine the partitioning of niches for NS5 species is a combination of abiotic conditions, such as temperature, and substrate utilisation. We propose that abiotic conditions influence spatial and temporal niche space across environments for each species, whereas substrate availability most strongly influences temporal niche dynamics within an environment.

We recognise that additional factors, particularly biotic interactions, can play an important role in determining a species' niche, but we were unable to address these in the scope of this study. Thus, further work would be required to understand the influence these factors have.

We present evidence here that NS5 genera are distinguishable by phylogeny, cell shape and size, genomic characteristics, spatiotemporal distribution patterns and predicted substrate metabolism. Based on this, we formally describe four novel candidate genera and type species within the *Flavobacteriaceae* family—etymology and metabolic descriptions are provided in Supplementary Material S1 and Supplementary Figs. S22–S25.

DATA AVAILABILITY

Metagenomes and MAGs used in this study were either previously deposited or deposited for this study in the European Nucleotide Archive, with all accession numbers provided in Supplementary Table S2. Those deposited for this study include the cultured isolate genome, PRJEB45371, and the MAGs derived from the South Pacific gyre metagenomes, PRJEB43746. The 16S rRNA gene amplicon time-series dataset used was previously published [63] and stored by JGI in the GOLD database under the project ID Gp0056779 as part of the community sequencing project COGITO. The ARB database containing the 16S rRNA gene NS5 phylogenetic tree is provided as Supplementary File S1. All data tables and the R script required to recreate the main body figures are available at https://github.com/tpriest0/NS5_marine_group_manuscript_figures.

REFERENCES

1. Hutchinson GE. Concluding remarks. Cold Spring Harb Symp Quant Biol. 1957;22:415–27.
2. Hutchinson GE. An introduction to population biology. New Haven, CT: Yale University Press; 1978.
3. Larkin AA, Martiny AC. Microdiversity shapes the traits, niche space, and biogeography of microbial taxa. Environ Microbiol Rep. 2017;9:55–70.
4. Mena C, Reglero P, Balbín R, Martín M, Santiago R, Sintes E. Seasonal niche partitioning of surface temperate open ocean prokaryotic communities. Front Microbiol. 2020;11:1749.
5. Sarmento H, Morana C, Gasol JM. Bacterioplankton niche partitioning in the use of phytoplankton-derived dissolved organic carbon: quantity is more important than quality. ISME J. 2016;10:2582–92.
6. Auladell A, Barberán A, Logares R, Garcés E, Gasol JM, Ferrera I. Seasonal niche differentiation among closely related marine bacteria. ISME J. 2022;16:178–89.
7. Avcı B, Krüger K, Fuchs BM, Teeling H, Amann RI. Polysaccharide niche partitioning of distinct *Polaribacter* clades during North Sea spring algal blooms. ISME J. 2020;14:1369–83.
8. Ghiglione J-F, Galand PE, Pommier T, Pedrós-Alió C, Maas EW, Bakker K, et al. Pole-to-pole biogeography of surface and deep marine bacterial communities. Proc Natl Acad Sci USA. 2012;109:17633–8.
9. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science. 2006;311:1737–40.
10. Wang Z, Juarez DL, Pan J-F, Blinebry SK, Gronniger J, Clark JS, et al. Microbial communities across nearshore to offshore coastal transects are primarily shaped by distance and temperature. Environ Microbiol. 2019;21:3862–72.
11. Herlemann DP, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. ISME J. 2011;5:1571–9.
12. Delmont TO, Kiefl E, Kilinc O, Esen OC, Uysal I, Rappé MS, et al. Single-amino acid variants reveal evolutionary processes that shape the biogeography of a global SAR11 subclade. eLife. 2019;8:e46497.
13. Teeling H, Fuchs BM, Bennke CM, Krüger K, Chafee M, Kappelmann L, et al. Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. eLife. 2016;5:e11888.
14. 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. 2012;336:608–11.
15. Alonso C, Warnecke F, Amann R, Pernthaler J. High local and global diversity of flavobacteria in marine plankton. Environ Microbiol. 2007;9:1253–66.
16. Ngugi DK, Stingl U. High-quality draft single-cell genome sequence of the NS5 Marine Group from the Coastal Red Sea. Genome Announc. 2018;26:e00565-18.
17. Meziti A, Kormas KA, Moustaka-Gouni M, Karayanni H. Spatially uniform but temporally variable bacterioplankton in a semi-enclosed coastal area. Syst Appl Microbiol. 2015;38:358–67.
18. Milici M, Vital M, Tomasch J, Badewien TH, Giebel H-A, Plumeier I, et al. Diversity and community composition of particle-associated and free-living bacteria in mesopelagic and bathypelagic Southern Ocean water masses: evidence of dispersal limitation in the Bransfield Strait. Limnol Oceanogr. 2017;62:1080–95.
19. Beman JM, Vargas SM, Vazquez S, Wilson JM, Yu A, Cairo A, et al. Biogeochemistry and hydrography shape microbial community assembly and activity in the eastern tropical North Pacific Ocean oxygen minimum zone. Environ Microbiol. 2020;23:2765–81.
20. Rapp JZ, Fernández-Méndez M, Bienhold C, Boetius A. Effects of ice-algal aggregate export on the connectivity of bacterial communities in the Central Arctic Ocean. Front Microbiol. 2018;9:01035.
21. Gómez-Pereira PR, Fuchs BM, Alonso C, Oliver MJ, van Beusekom JEE, Amann R. Distinct flavobacterial communities in contrasting water masses of the North Atlantic Ocean. ISME J. 2010;4:472–87.
22. Choi DH, An SM, Yang EC, Lee H, Shim J, Jeong J, et al. Daily variation in the prokaryotic community during a spring bloom in shelf waters of the East China Sea. FEMS Microbiol Ecol. 2018;94:fiy134.
23. Yang C, Li Y, Zhou B, Zhou Y, Zheng W, Tian Y, et al. Illumina sequencing-based analysis of free-living bacterial community dynamics during an Akashiwo sanguine bloom in Xiamen sea, China. Sci Rep. 2015;5:8476.

24. Díez-Vives C, Nielsen S, Sánchez P, Palenzuela O, Ferrera I, Sebastián M, et al. Delineation of ecologically distinct units of marine *Bacteroidetes* in the Northwestern Mediterranean Sea. Mol Ecol. 2019;28:2846–59.
25. Seo J-H, Kang I, Yang S-J, Cho J-C. Characterization of spatial distribution of the bacterial community in the South Sea of Korea. PLoS ONE. 2017;12:e0174159.
26. Alonso-Sáez L, Díaz-Pérez L, Morán XAG. The hidden seasonality of the rare biosphere in coastal marine bacterioplankton. Environ Microbiol. 2015;17:3766–80.
27. Priest T, Orellana LH, Huettel B, Fuchs BM, Amann R. Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms. PeerJ. 2021;9:e11721.
28. Zhou J, Bruns MA, Tiedje JM. DNA recovery from soils of diverse composition. Appl Environ Microbiol. 1996;62:316–22.
29. Li D, Luo R, Liu C-M, Leung C-M, Ting H-F, Sadakane K, et al. MEGAHIT v1.0: a fast and scalable metagenome assembler driven by advanced methodologies and community practices. Methods. 2016;102:3–11.
30. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. Nat Methods. 2020;17:1103–10.
31. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. Nat Methods. 2014;11:1144–6.
32. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. PeerJ. 2019;7:e7359.
33. Wu Y-W, Simmons BA, Singer SW. MaxBin 2.0: an automated binning algorithm to recover genomes from multiple metagenomic datasets. Bioinformatics. 2016;32:605–7.
34. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. Nat Microbiol. 2018;3:836–43.
35. 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;25:1043–55.
36. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. Bioinformatics. 2020;36:1925–7.
37. Parks DH, Chuvochina M, Chaumeil P-A, Rinke C, Mussig AJ, Hugenholtz P. A complete domain-to-species taxonomy for Bacteria and Archaea. Nat Biotechnol. 2020;38:1079–86.
38. Krüger K, Chafee M, Ben Francis T, Glavina del Rio T, Becher D, Schweder T, et al. In marine *Bacteroidetes* the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. ISME J. 2019;13:2800–16.
39. Francis TB, Bartosik D, Sura T, Sichert A, Hehemann J-H, Markert S, et al. Changing expression patterns of TonB-dependent transporters suggest shifts in polysaccharide consumption over the course of a spring phytoplankton bloom. ISME J. 2021;15:2336–50.
40. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. Nat Commun. 2018;9:5114.
41. Winkelmann N, Harder J. An improved isolation method for attached-living *Planctomycetes* of the genus *Rhodopirellula*. J Microbiol Methods. 2009;77:276–84.
42. Hahnke RL, Bennke CM, Fuchs BM, Mann AJ, Rhiel E, Teeling H, et al. Dilution cultivation of marine heterotrophic bacteria abundant after a spring phytoplankton bloom in the North Sea. Environ Microbiol. 2015;17:3515–26.
43. Koren S, Walenz BP, Berlin K, Miller JR, Bergman NH, Phillippy AM. Canu: scalable and accurate long-read assembly via adaptive k-mer weighting and repeat separation. Genome Res. 2017;27:722–36.
44. 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. J Comput Biol. 2012;19:455–77.
45. 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;1:1–6.
46. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res. 2004;32:1792–7.
47. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. Bioinformatics. 2009;25:1972–3.
48. Price MN, Dehal PS, Arkin AP. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE. 2010;5:e9490.
49. Seeman T. Barrnap 0.9 (version 3): rapid ribosomal RNA prediction. 2017. https://github.com/tseemann/barrnap.
50. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: a software environment for sequence data. Nucleic Acids Res. 2004;32:1363–71.
51. Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. Bioinformatics. 2012;28:1823–9.
52. Stamatakis A. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. Bioinformatics. 2014;30:1312–3.
53. Amann RI, Krumholz L, Stahl DA. Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. J Bacteriol. 1990;172:762–70.
54. Pernthaler A, Pernthaler J, Amann R. Fluorescence in situ hybridization and catalyzed reporter deposition for the identification of marine bacteria. Appl Environ Microbiol. 2002;68:3094–101.
55. Pesant S, Not F, Picheral M, Kandels-Lewis S, Le Bescot N, Gorsky G, et al. Open science resources for the discovery and analysis of Tara Oceans data. Sci Data. 2015;2:150023.
56. Bushnell B. BBTools software package. 2017. https://sourceforge.net/projects/bbmap/.
57. Mortazavi A, Williams BA, McCue K, Schaeffer L, Wold B. Mapping and quantifying mammalian transcriptomes by RNA-Seq. Nat Methods. 2008;5:621–8.
58. RStudio Team. RStudio: integrated development of R. Boston, MA: RStudio Inc.; 2015.
59. South A. rnaturalearth: World map data from Natural Earth. R packag version 0.1.0; 2017.
60. Pebesma E. Simple features for R: standardized support for spatial vector data. R J. 2018;10:439–46.
61. Wickham H. ggplot2: elegant graphics for data analysis. New York: Springer-Verlag; 2016.
62. Orellana LH, Francis TB, Ferraro M, Hehemann J-H, Fuchs BM, Amann RI. *Verrucomicrobiota* are specialist consumers of sulfated methyl pentoses during diatom blooms. ISME J. 2021.
63. Chafee M, Fernàndez-Guerra A, Buttigieg PL, Gerdts G, Eren AM, Teeling H, et al. Recurrent patterns of microdiversity in a temperate coastal marine environment. ISME J. 2018;12:237–52.
64. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. Vegan community ecology package version 2.5, 7 November. 2020.
65. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. Bioinformatics. 2020;36:2251–2.
66. Overbeek R, Olson R, Pusch GD, Olsen GJ, Davis JJ, Disz T, et al. The SEED and the rapid annotation of microbial genomes using subsystems technology (RAST). Nucleic Acids Res. 2014;42:D206–14.
67. Seemann T. Prokka: rapid prokaryotic genome annotation. Bioinformatics. 2014;30:2068–9.
68. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. Nucleic Acids Res. 2018;46:95–101.
69. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. Nat Methods. 2015;12:59–60.
70. Lombard V, Golaconda Ramulu H, Drula E, Coutinho PM, Henrissat B. The carbohydrate-active enzymes database (CAZy) in 2013. Nucleic Acids Res. 2014;42:490–5.
71. Barbeyron T, Brillet-Guéguen L, Carré W, Carrière C, Caron C, Czjzek M, et al. Matching the diversity of sulfated biomolecules: creation of a classification database for sulfatases reflecting their substrate specificity. PLoS ONE. 2016;11:e0164846.
72. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. Nucleic Acids Res. 2018;46:624–32.
73. Wilkins D. gggenes: Draw Gene Arrow Maps in 'ggplot2'. R package version 0.4.1; 2020.
74. de Vries A, Ripley BD. ggdendro: create dendrograms and tree diagrams using 'ggplot2'. 2020.
75. Kappelmann L, Krüger K, Hehemann J-H, Harder J, Markert S, Unfried F, et al. Polysaccharide utilization loci of North Sea *Flavobacteriia* as basis for using SusC/D-protein expression for predicting major phytoplankton glycans. ISME J. 2019;13:76–91.
76. Katoh K, Misawa K, Kuma K, Miyata T. MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. Nucleic Acids Res. 2002;30:3059–66.
77. Letunic I, Bork P. Interactive tree of life (iTOL) v4: recent updates and new developments. Nucleic Acids Res. 2019;47:256–9.
78. Yarza P, Yilmaz P, Pruesse E, Glöckner FO, Ludwig W, Schleifer K-H, et al. Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. Nat Rev Microbiol. 2014;12:635–45.
79. Konstantinidis KT, Rosselló-Móra R, Amann R. Uncultivated microbes in need of their own taxonomy. ISME J. 2017;11:2399–406.
80. Bjursell MK, Martens EC, Gordon JI. Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, *Bacteroides thetaiotaomicron*, to the suckling period. J Biol Chem. 2006;281:36269–79.

81. Ficko-Blean E, Préchoux A, Thomas F, Rochat T, Larocque R, Zhu Y, et al. Carrageenan catabolism is encoded by a complex regulon in marine heterotrophic bacteria. Nat Commun. 2017;8:1685.
82. Johnson ET, Baron DB, Naranjo B, Bond DR, Schmidt-Dannert C, Gralnick JA. Enhancement of survival and electricity production in an engineered bacterium by light-driven proton pumping. Appl Environ Microbiol. 2010;76:4123–9.
83. Dubinsky V, Haber M, Burgsdorf I, Saurav K, Lehahn Y, Malik A, et al. Metagenomic analysis reveals unusually high incidence of proteorhodopsin genes in the ultraoligotrophic Eastern Mediterranean Sea. Environ Microbiol. 2017;19:1077–90.
84. Fernández-Gómez B, Richter M, Schüler M, Pinhassi J, Acinas SG, González JM, et al. Ecology of marine *Bacteroidetes*: a comparative genomics approach. ISME J. 2013;7:1026–37.
85. Heins A, Reintjes G, Amann RI, Harder J. Particle collection in Imhoff sedimentation cones enriches both motile chemotactic and particle-attached bacteria. Front Microbiol. 2021;12:643730.
86. Unfried F, Becker S, Robb CS, Hehemann J-H, Markert S, Heiden SE, et al. Adaptive mechanisms that provide competitive advantages to marine *Bacteroidetes* during microalgal blooms. ISME J. 2018;12:2894–906.
87. Bauer M, Kube M, Teeling H, Richter M, Lombardot T, Allers E, et al. Whole genome analysis of the marine *Bacteroidetes* '*Gramella forsetii*' reveals adaptations to degradation of polymeric organic matter. Environ Microbiol. 2006;8:2201–13.
88. Kabisch A, Otto A, König S, Becher D, Albrecht D, Schüler M, et al. Functional characterization of polysaccharide utilization loci in the marine *Bacteroidetes* '*Gramella forsetii*' KT0803. ISME J. 2014;8:1492–502.
89. Reintjes G, Arnosti C, Fuchs B, Amann R. Selfish, sharing and scavenging bacteria in the Atlantic Ocean: a biogeographical study of bacterial substrate utilisation. ISME J. 2019;13:1119–32.
90. Thomas F, Barbeyron T, Tonon T, Génicot S, Czjzek M, Michel G. Characterization of the first alginolytic operons in a marine bacterium: from their emergence in marine *Flavobacteriia* to their independent transfers to marine *Proteobacteria* and human gut *Bacteroides*. Environ Microbiol. 2012;14:2379–94.
91. Hehemann J-H, Arevalo P, Datta MS, Yu X, Corzett CH, Henschel A, et al. Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. Nat Commun. 2016;7:12860.
92. Deniaud-Bouët E, Hardouin K, Potin P, Kloareg B, Hervé C. A review about brown algal cell walls and fucose-containing sulfated polysaccharides: cell wall context, biomedical properties and key research challenges. Carbohydr Polym. 2017;175:395–408.
93. 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;5:1026–39.
94. Duerschlag J, Mohr W, Ferdelman TG, LaRoche J, Desai D, Croot PL, et al. Niche partitioning by photosynthetic plankton as a driver of CO2-fixation across the oligotrophic South Pacific Subtropical Ocean. ISME J. 2022;15:465–76.
95. Sunagawa S, Coelho LP, Chaffron S, Kultima JR, Labadie K, Salazar G, et al. Structure and function of the global ocean microbiome. Science. 2015;348:1261359.
96. Raes J, Letunic I, Yamada T, Jensen LJ, Bork P. Toward molecular trait-based ecology through integration of biogeochemical, geographical and metagenomic data. Mol Syst Biol. 2011;7:473.
97. Baas Becking L. G. M.. Geobiologie of inleiding tot de milieukunde. WP Van Stock Zoon, Den Haag; 1934.
98. Gibbons SM, Caporaso JG, Pirrung M, Field D, Knight R, Gilbert JA. Evidence for a persistent microbial seed bank throughout the global ocean. Proc Natl Acad Sci USA. 2013;110:4651–5.
99. Lennon JT, Jones SE. Microbial seed banks: the ecological and evolutionary implications of dormancy. Nat Rev Microbiol. 2011;9:119–30.

ACKNOWLEDGEMENTS

We thank Sabine Kühn for her technical support on isolation and cultivation. We thank Monike Oggerin for providing the metagenome-assembled genomes from the South Pacific Gyre dataset. We thank Bruno Huettel and the entire team at the Max-Planck-Genome-centre Cologne (http://mpgc.mpipz.mpg.de/home/) for their efforts with genome and metagenome sequencing. We thank Aharon Oren for his advice on etymology. We thank Susanne Erdmann for her assistance with the transmission electron microscopy. TP is a member of the International Max Planck Research School of Marine Microbiology (MarMic). This study was funded by the Max Planck Society.

AUTHOR CONTRIBUTIONS

TP, BMF and RA conceived and designed the study. TP performed all bioinformatic and molecular analyses. AH and JH performed the isolation and cultivation of the isolate. TP wrote the manuscript with contributions and input from all coauthors. All authors read and approved the final version of the manuscript.

FUNDING

Open Access funding enabled and organized by Projekt DEAL.

COMPETING INTERESTS

The authors declare no competing interests.

ADDITIONAL INFORMATION

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41396-022-01209-8.

Correspondence and requests for materials should be addressed to Bernhard M. Fuchs.

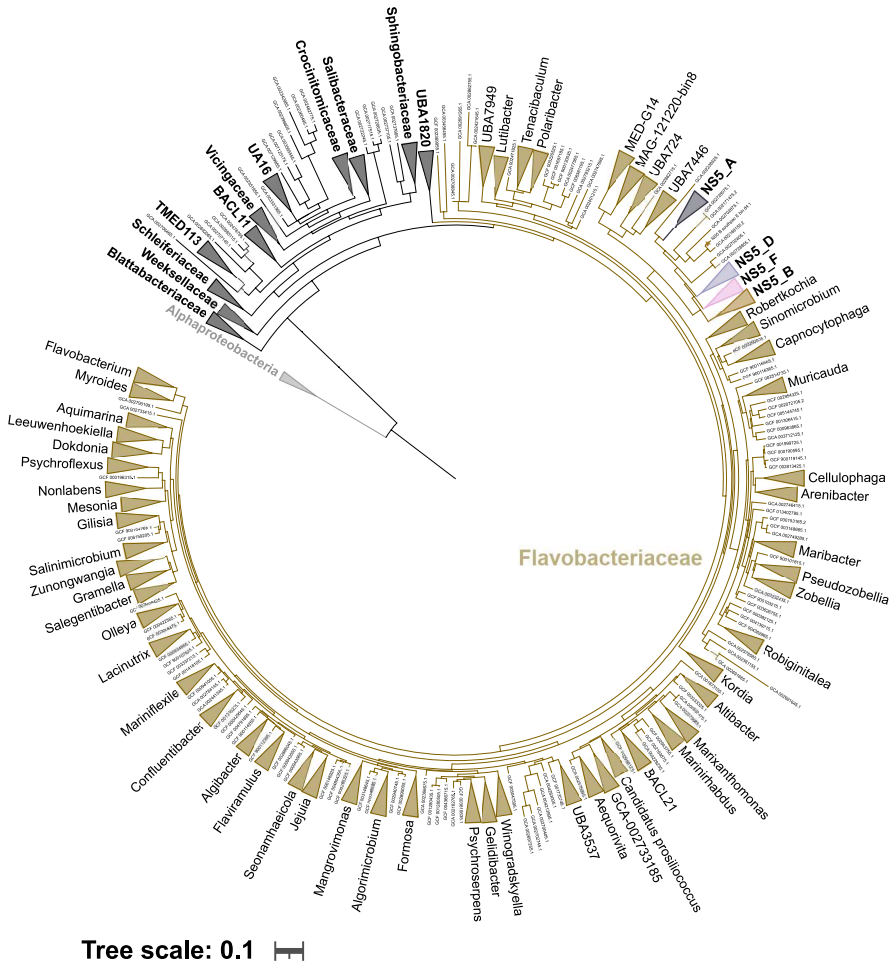
Reprints and permission information is available at http://www.nature.com/reprints

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Supplementary tables and figures

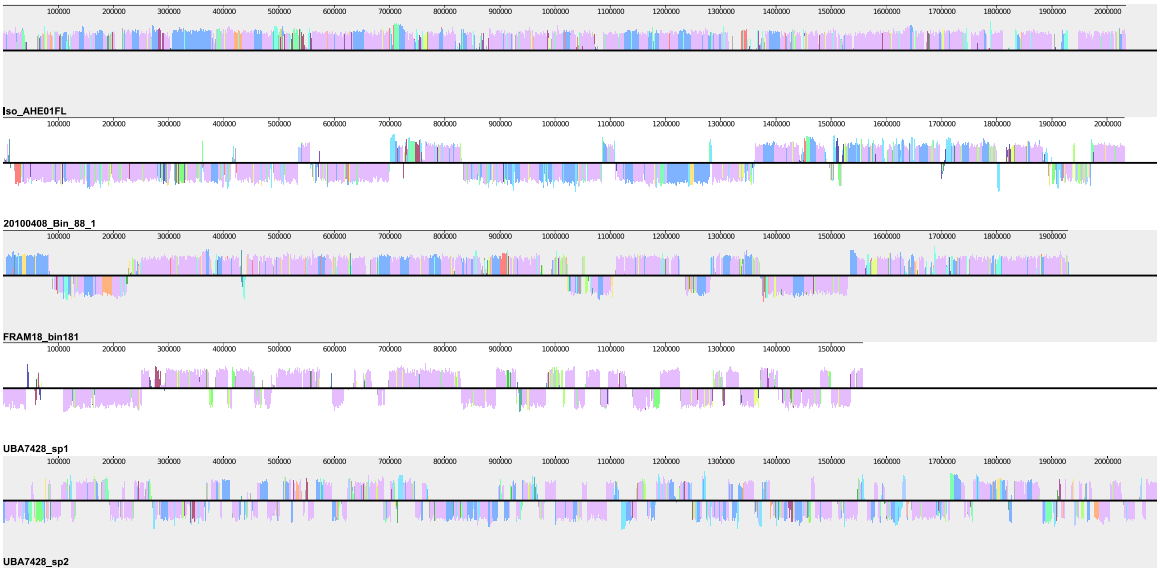
All supplementary tables are available on the USB attached with this thesis and at the following link. Due to their large size, they were not included in the printed thesis.

https://owncloud.mpi-bremen.de/index.php/s/bYyADuyJyIE1xup

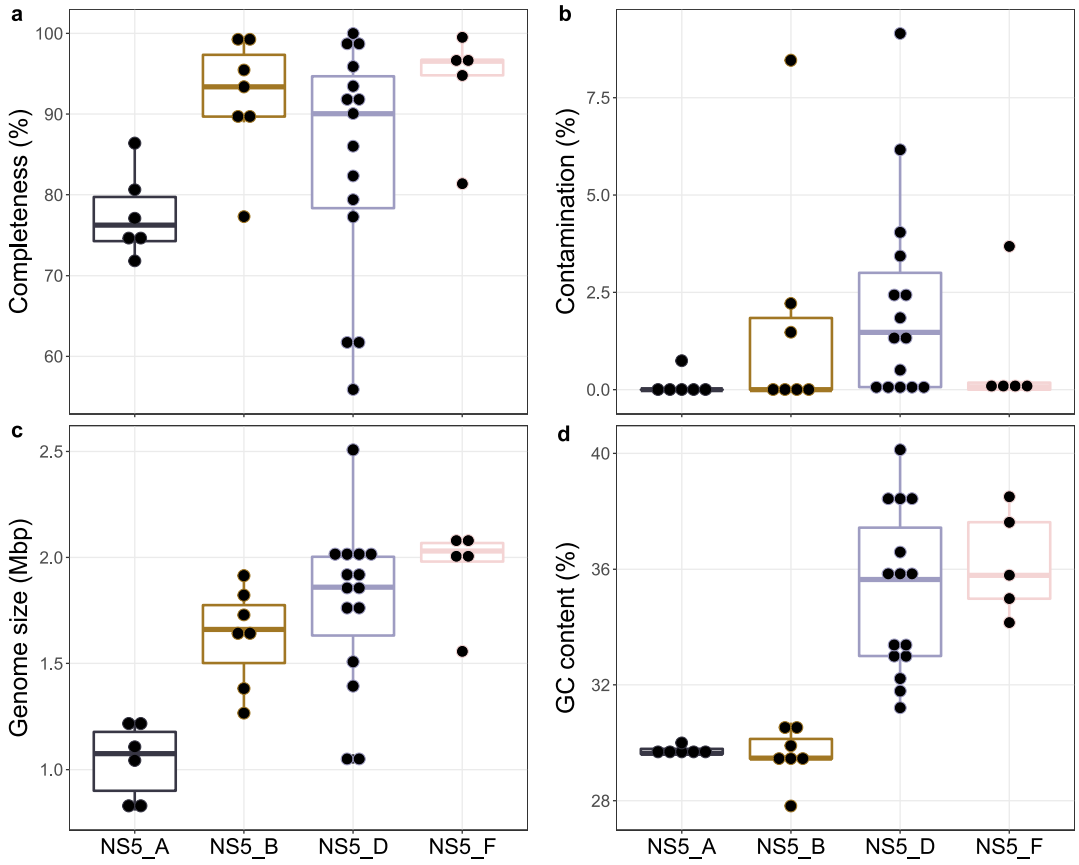


Tree scale: 0.1

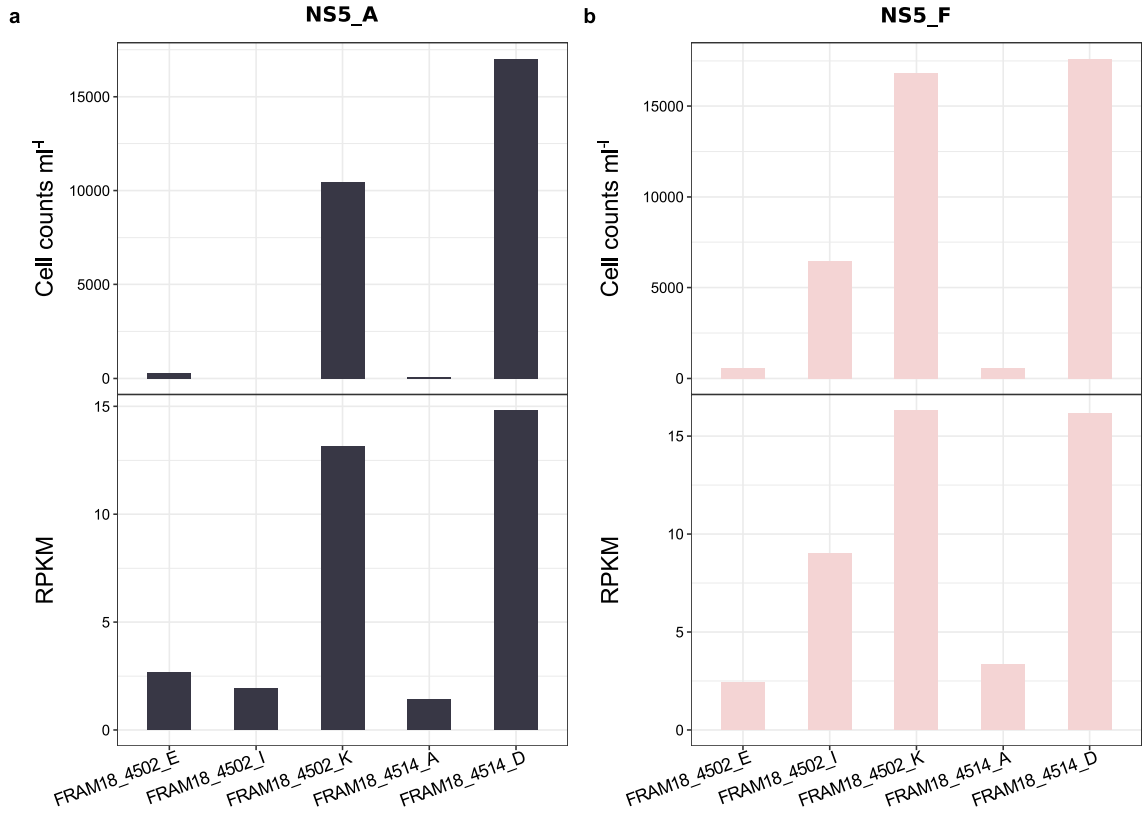
Supplementary Figure S1. Ribosomal protein tree of the Flavobacteriia class. The tree was constructed using a concatenated alignment of 16 ribosomal proteins from 1275 complete genome assemblies from RefSeq along with the 35 species-representaive NS5 MAG sequences. One species from NS5_B, indicated with a *, was positioned outside of the gens-level cluster.



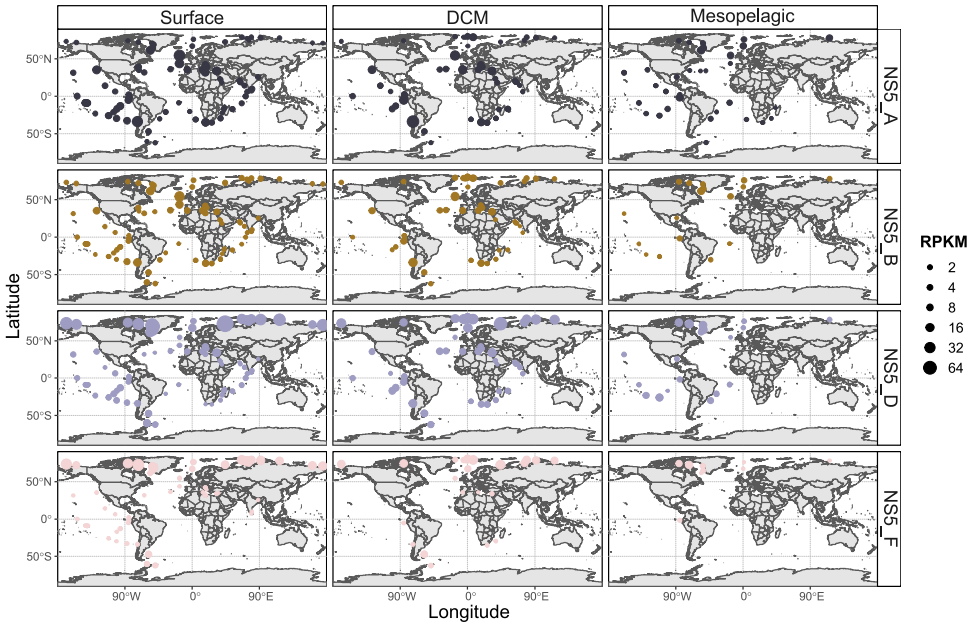
Supplementary Figure S2. Genome alignment and conserved syntenic block identification for species-representatives in NS5_F against the isolate genome, Iso_AHE01FL. Alignment was performed using the progressiveMauve aligner in the Mauve program with default settings. The colours represent conserved syntenic gene blocks, with the same colour across different genomes indicating a shared block. The regions shared by all speices-representatives are coloured in mauve.



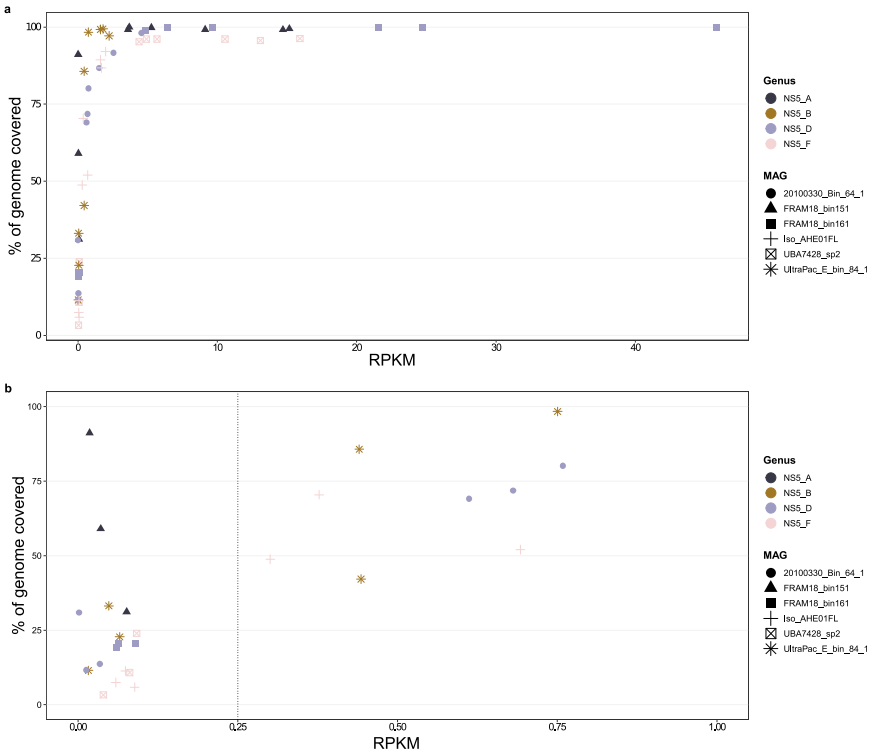
Supplementary Figure S3. Summary statistics of species-representative MAGs. All values were determined using CheckM.



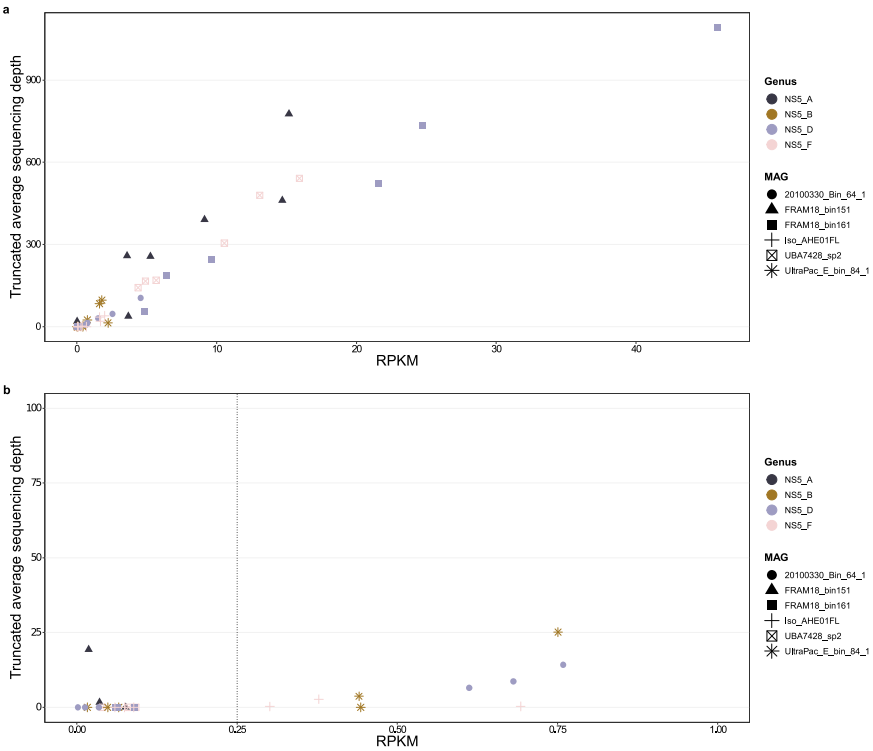
Supplementary Figure S4. Comparison of absolute cell counts and RPKM values from surface seawater samples, 0.2 - 3 µm fraction, taken in the Fram Strait region. Cell counts were derived from CARD-FISH analysis, where cells containing the group-specific probe signal and a nucleic acid stain were enumerated. RPKM values were based on read recruitment from sample metagenomes using BBMap with a 99% identity threshold. Samples used for this analysis were derived from [26], with sample names being maintained for comparison.



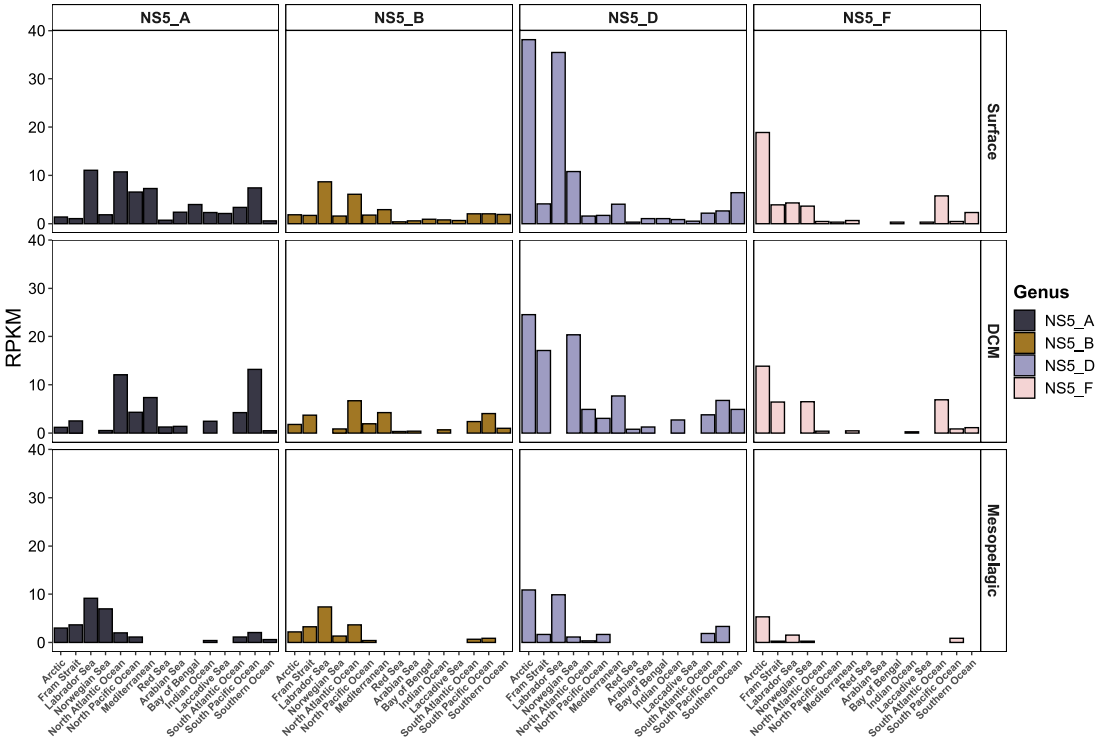
Supplementary Figure S5. Global distribution of NS5 genera from the surface to mesopelagic layer. Distribution was determined by read recruitment from TaraOceans metagenomic samples to each species-representative MAG using BBMap with a 99% identity threshold cut-off. The number of mapped reads was converted to RPKM values and a minimum threshold of 0.25 RPKM was applied, which ensured a minimum coverage of 40%. Genus RPKM values were obtained by summing the respective species' RPKM values for each sample.



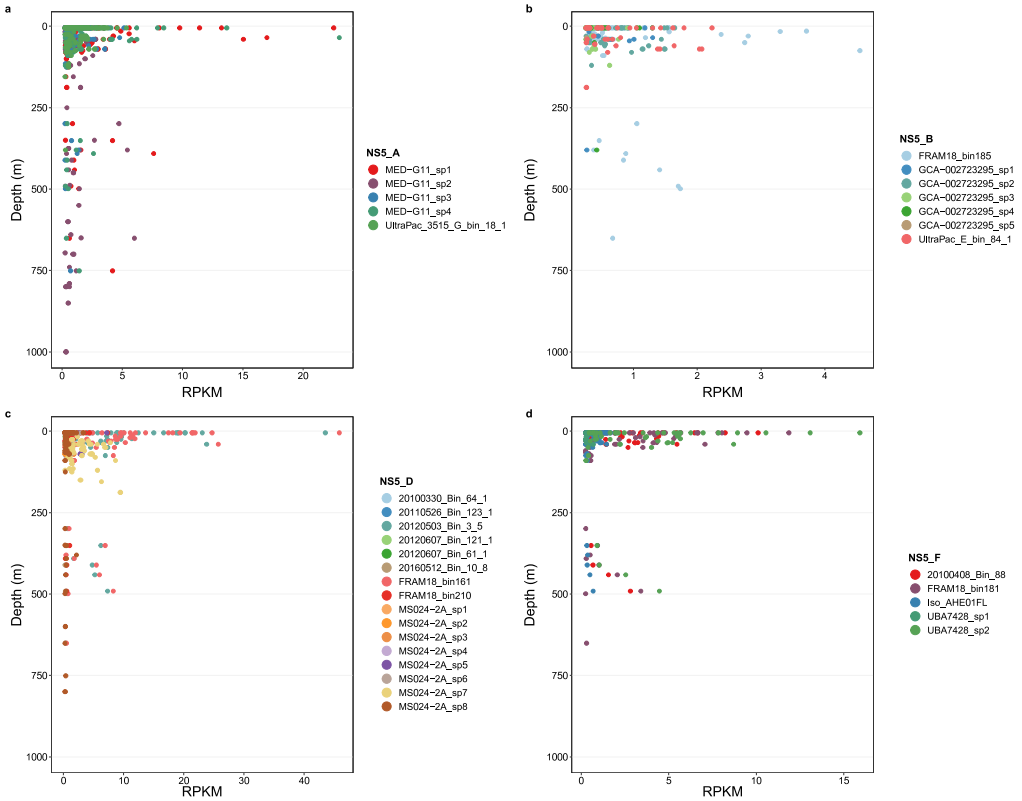
Supplementary Figure S6. RPKM values compared to genome coverage from the mapping of reads from Tara Oceans metagenome samples to selected NS5 species representatives. Representatives were selected based on exhibiting a large range in RPKM values across samples, and the samples in which they exhibited high, medium and low RPKM values were chosen for visualisation in this figure. Read recruitment was performed using BBMap with an identity threshold of 99%.



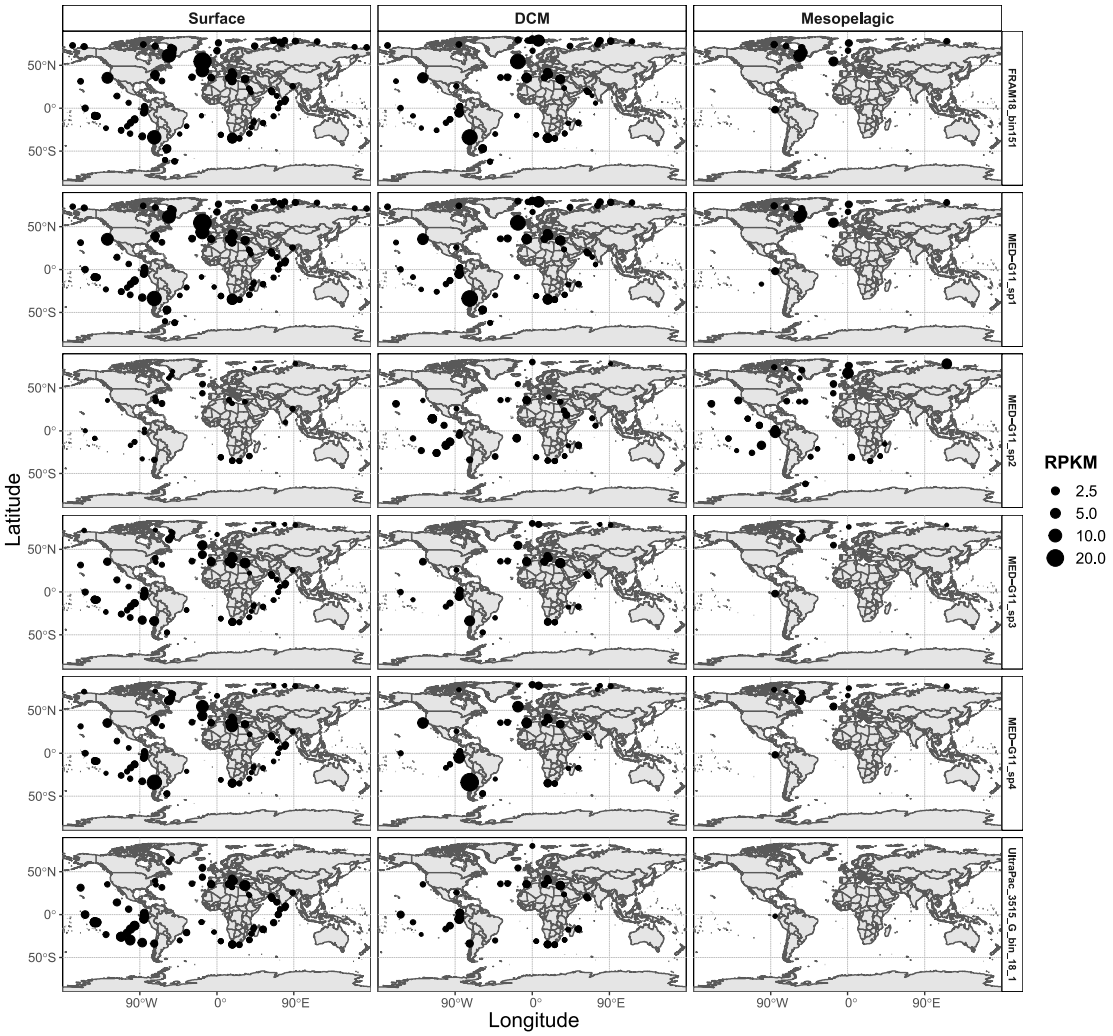
Supplementary Figure S7. RPKM values compared to truncated average depth from the mapping of reads from Tara Oceans metagenome samples to selected NS5 species representatives. Representatives were selected based on exhibiting a large range in RPKM values across samples, and the samples in which they exhibited high, medium and low RPKM values were chosen for visualisation in this figure. Read recruitment was performed using BBMap with an identity threshold of 99%. Truncated average depth values were calculated according to [61].



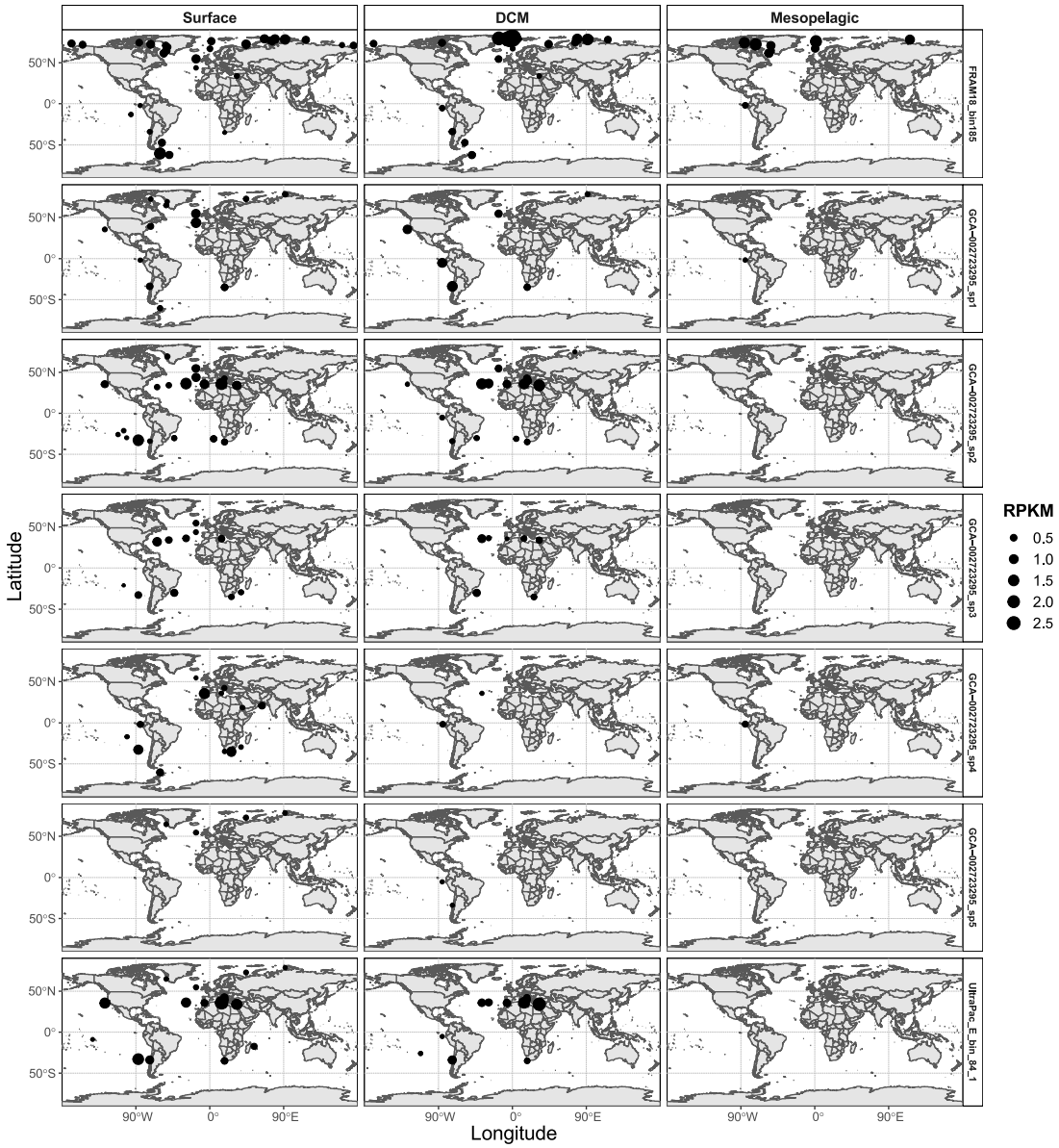
Supplementary Figure S8. Distribution of NS5 genera across oceanic regions. Reads were recruited from Tara Oceans metagenomes to each species-representative MAG using BBMap with a 99% identity threshold cut-off and subsequently converted to RPKM values. Genus-level RPKM values were derived from the sum of the respective species' RPKM values in each sample. Tara Oceans metagenomes were designated into oceanic regions based on latitude and longitude. RPKM value for each genus in each region was derived from the average RPKM values of the genus in all samples of that region.



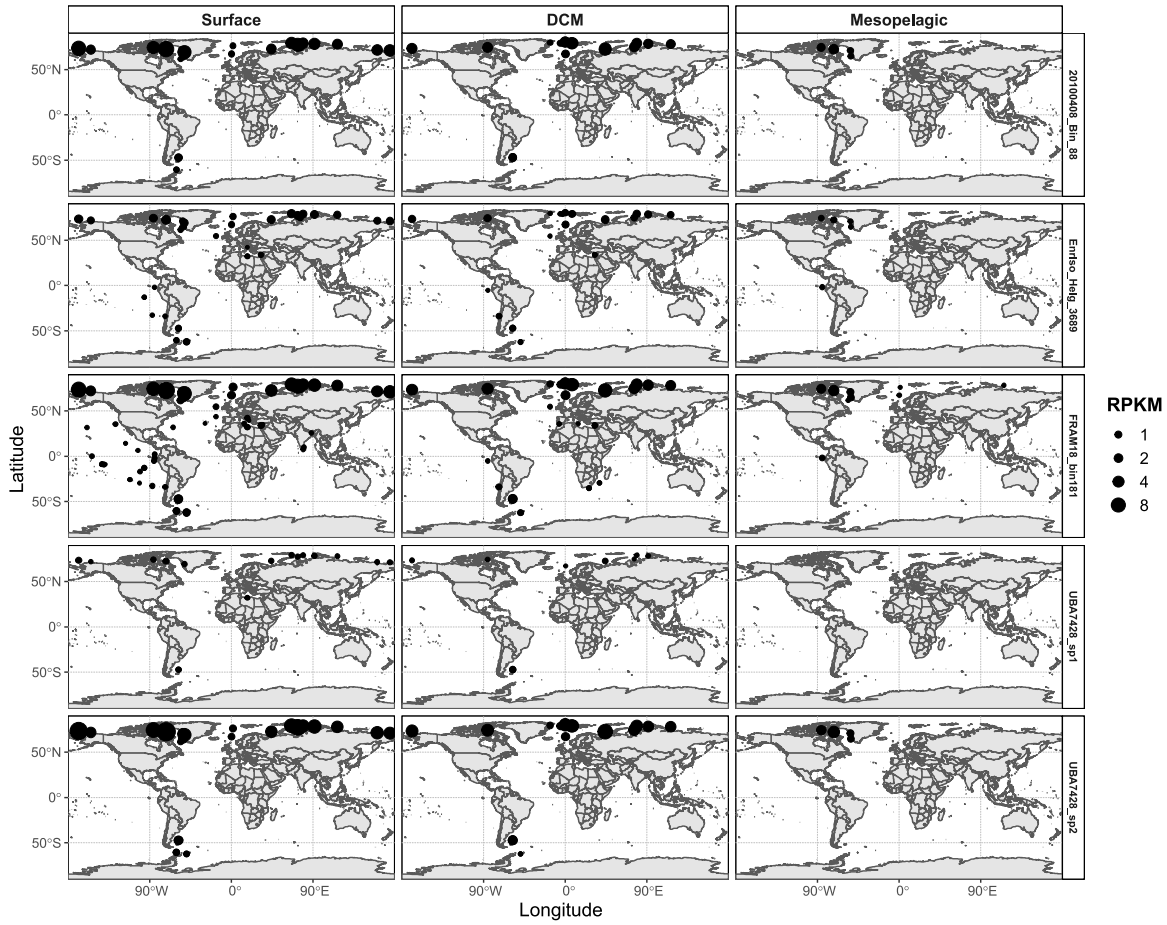
Supplementary Figure S9. RPKM values of NS5 species representatives in relation to depth across Tara Oceans samples. RPKM values were determined through read recruitment from Tara Oceans metagenomes using BBMap with a 99% identity threshold.



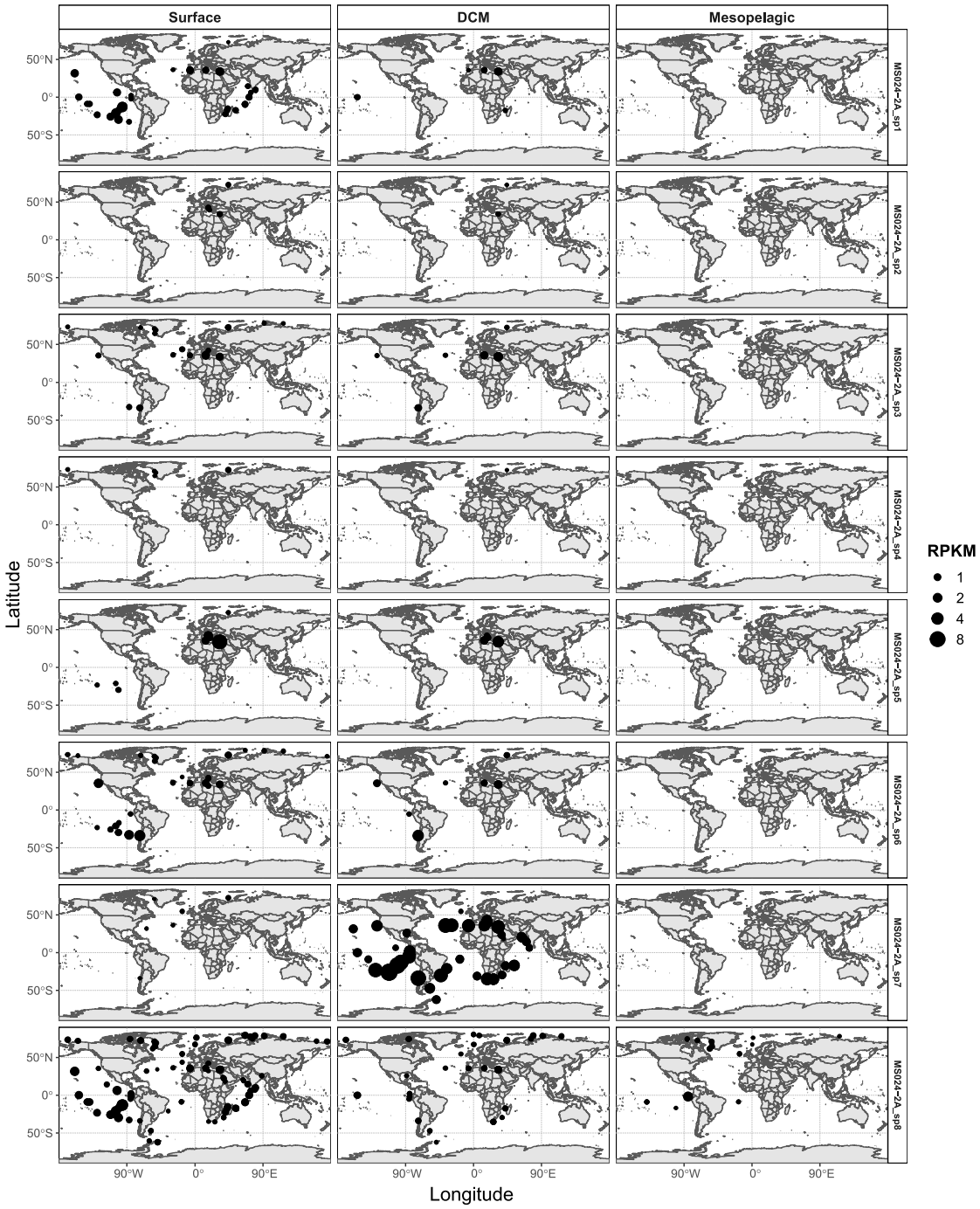
Supplementary Figure S10. Global distribution of the species-representatives within NS5_A from the surface to mesopelagic layer. RPKM values were determined from read recruitment of Tara Oceans metagenomic reads to each representative using BBMap with a 99% identity threshold cut-off. Resulting RPKM values <0.25 were excluded from the plot, which ensured a minimum genome coverage of 40%.



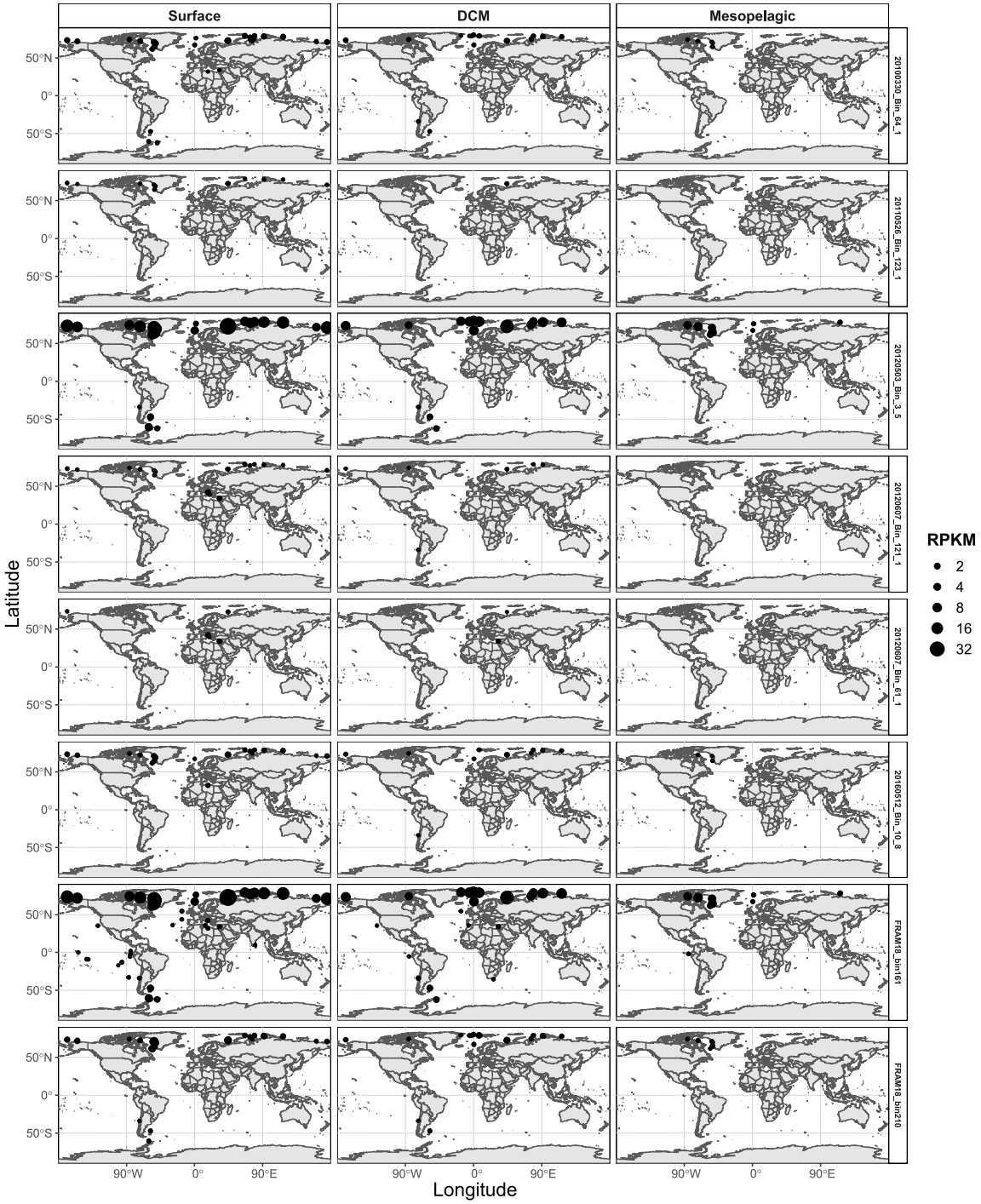
Supplementary Figure S11. Global distribution of the species-representatives within NS5_B from the surface to mesopelagic layer. RPKM values were determined from read recruitment of Tara Oceans metagenomic reads to each representative using BBMap with a 99% identity threshold cut-off. Resulting RPKM values <0.25 were excluded from the plot, which ensured a minimum genome coverage of 40%.



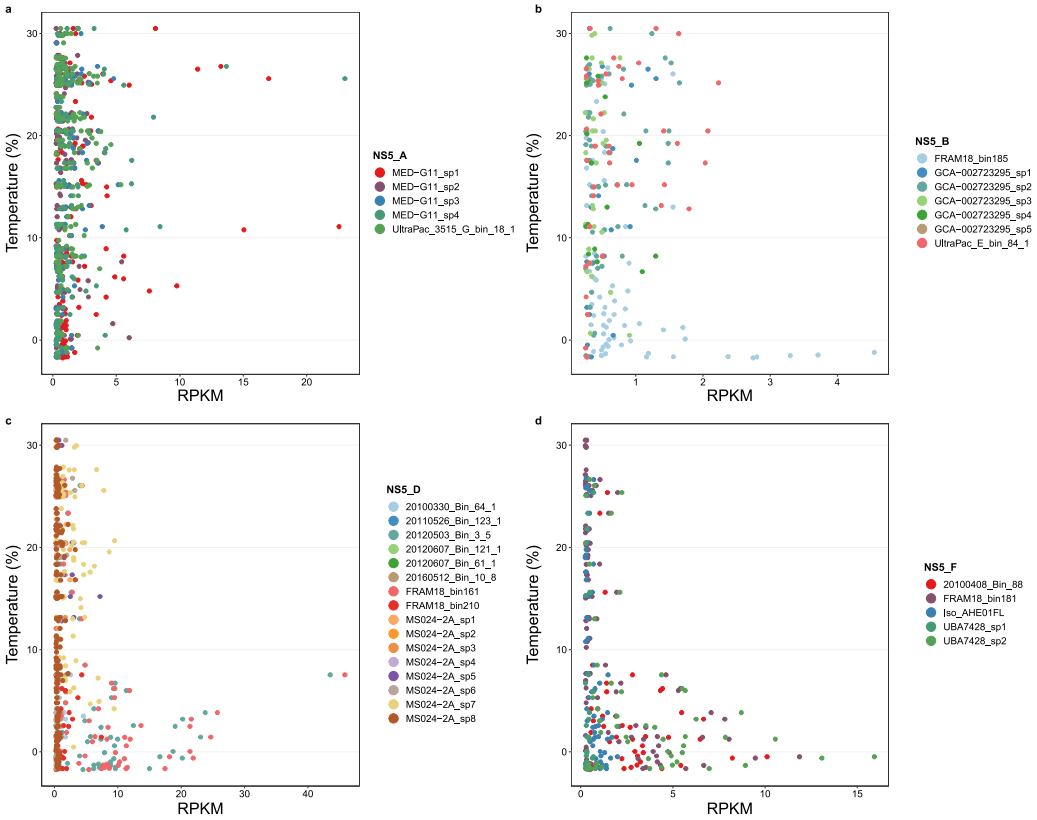
Supplementary Figure S12. Global distribution of the species-representatives within NS5_F from the surface to mesopelagic layer. RPKM values were determined from read recruitment of Tara Oceans metagenomic reads to each representative using BBMap with a 99% identity threshold cut-off. Resulting RPKM values <0.25 were excluded from the plot, which ensured a minimum genome coverage of 40%.



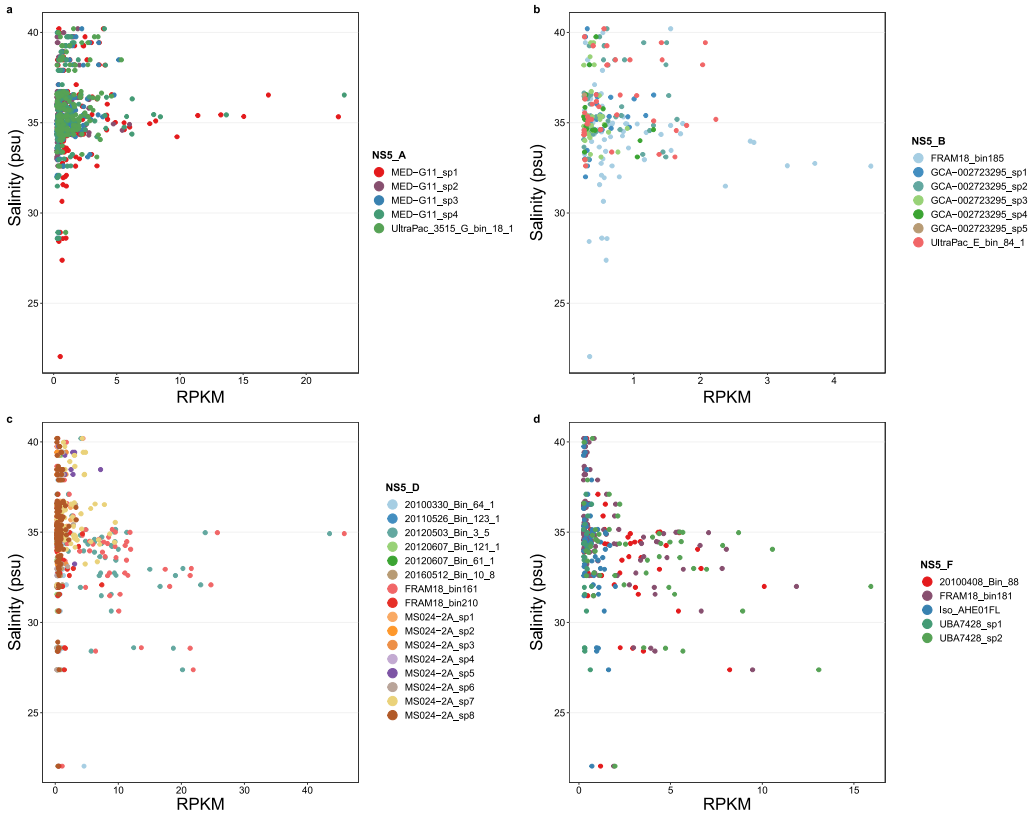
Supplementary Figure S13. Global distribution of the species-representatives within NS5_D from the surface to mesopelagic layer. RPKM values were determined from read recruitment of Tara Oceans metagenomic reads to each representative using BBMap with a 99% identity threshold cut-off. Resulting RPKM values <0.25 were excluded from the plot, which ensured a minimum genome coverage of 40%. Due to the large number of species in NS5_D, this plot includes only half, whilst the other half are visualised in Supplementary Figure S14.



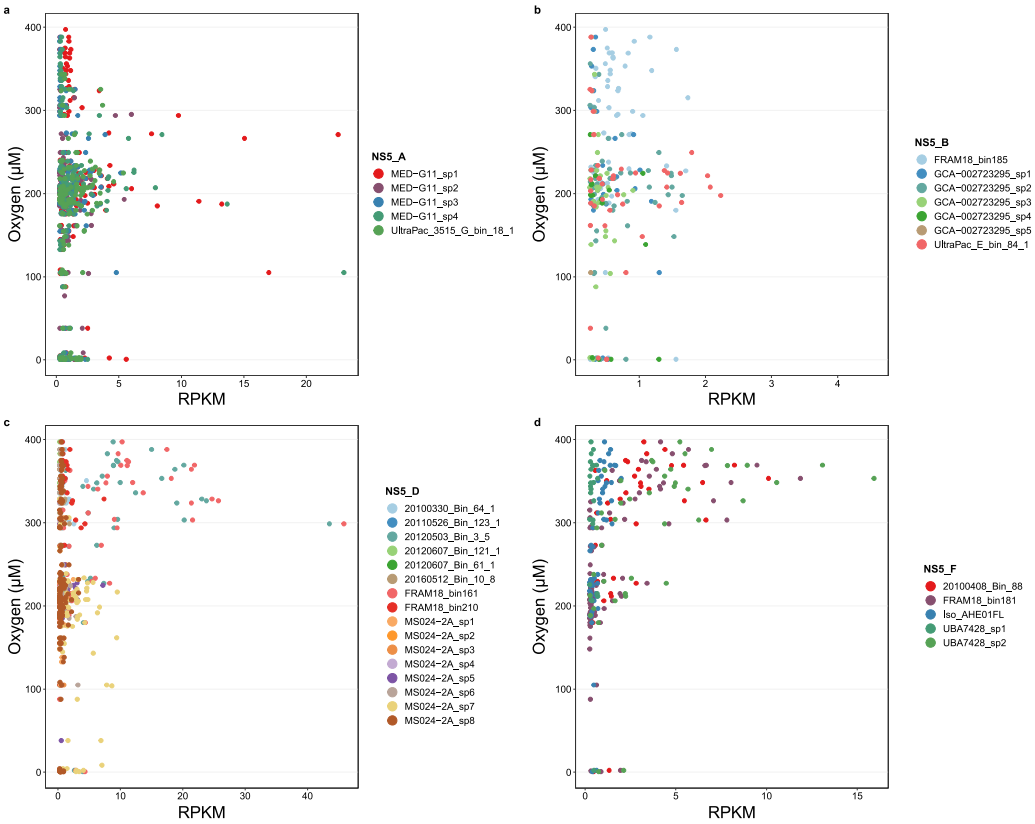
Supplementary Figure S14. Global distribution of the species-representatives within NS5_D from the surface to mesopelagic layer. RPKM values were determined from read recruitment of Tara Oceans metagenomic reads to each representative using BBMap with a 99% identity threshold cut-off. Resulting RPKM values <0.25 were excluded from the plot, which ensured a minimum genome coverage of 40%. Due to the large number of species in NS5_D, this plot includes only half, whilst the other half are visualised in Supplementary Figure S13.



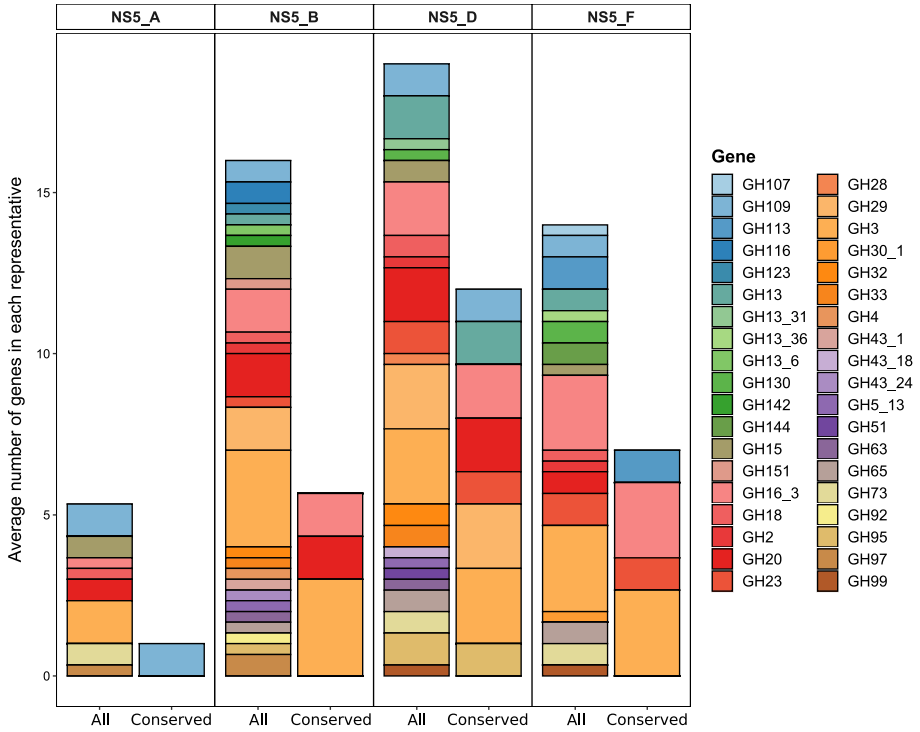
Supplementary Figure S15. RPKM values of NS5 species representatives in relation to temperature across Tara Oceans samples. RPKM values were determined through read recruitment from Tara Oceans metagenomes using BBMap with a 99% identity threshold.



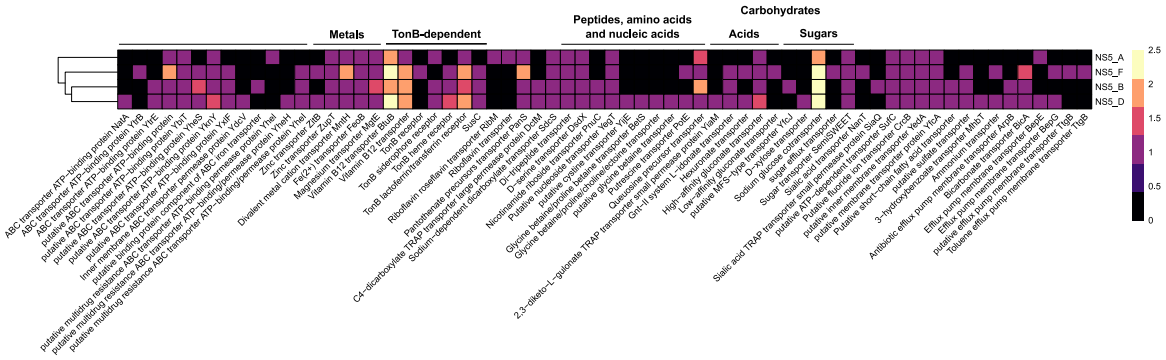
Supplementary Figure S16. RPKM values of NS5 species representatives in relation to salinity across Tara Oceans samples. RPKM values were determined through read recruitment from Tara Oceans metagenomes using BBMap with a 99% identity threshold.



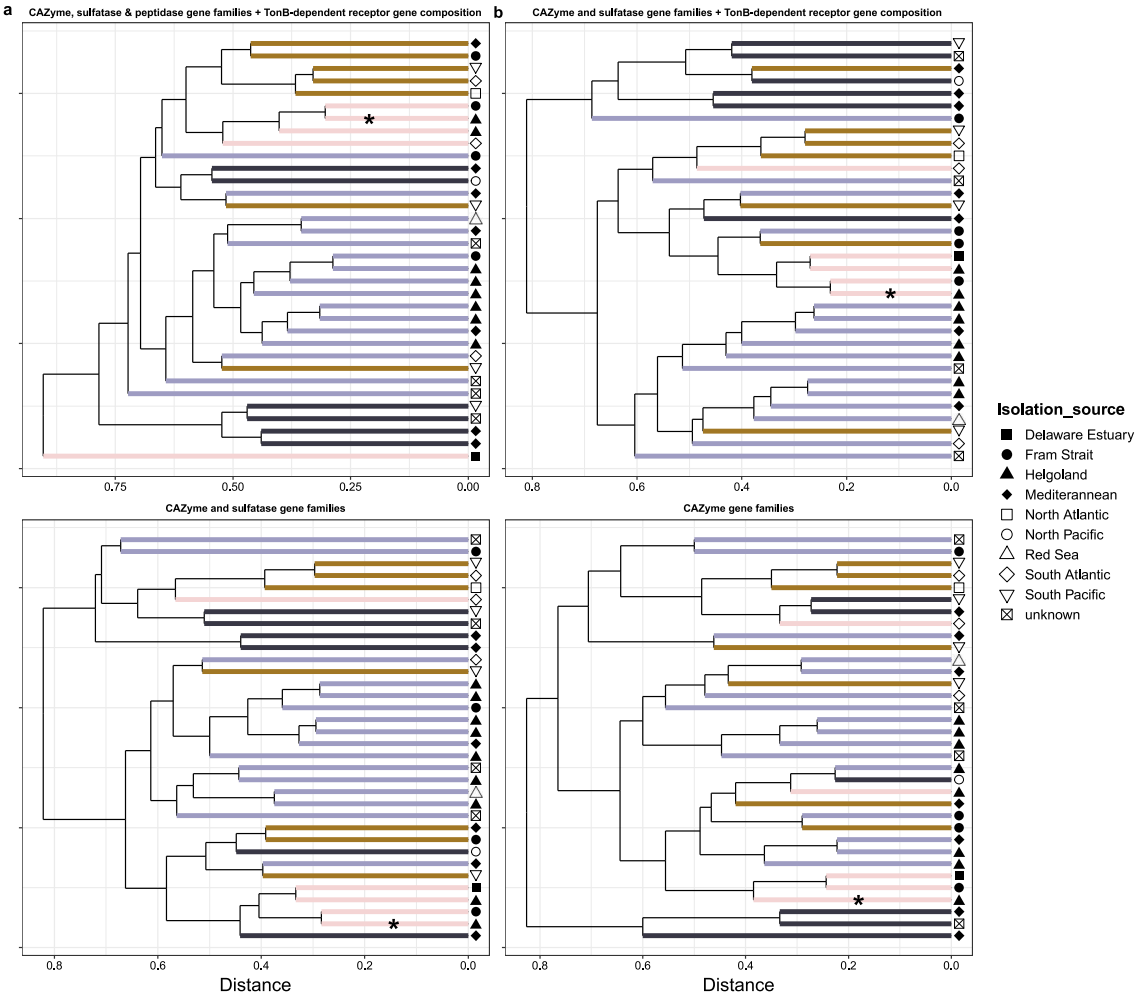
Supplementary Figure S17. RPKM values of NS5 species representatives in relation to oxygen across Tara Oceans samples. RPKM values were determined through read recruitment from Tara Oceans metagenomes using BBMap with a 99% identity threshold.



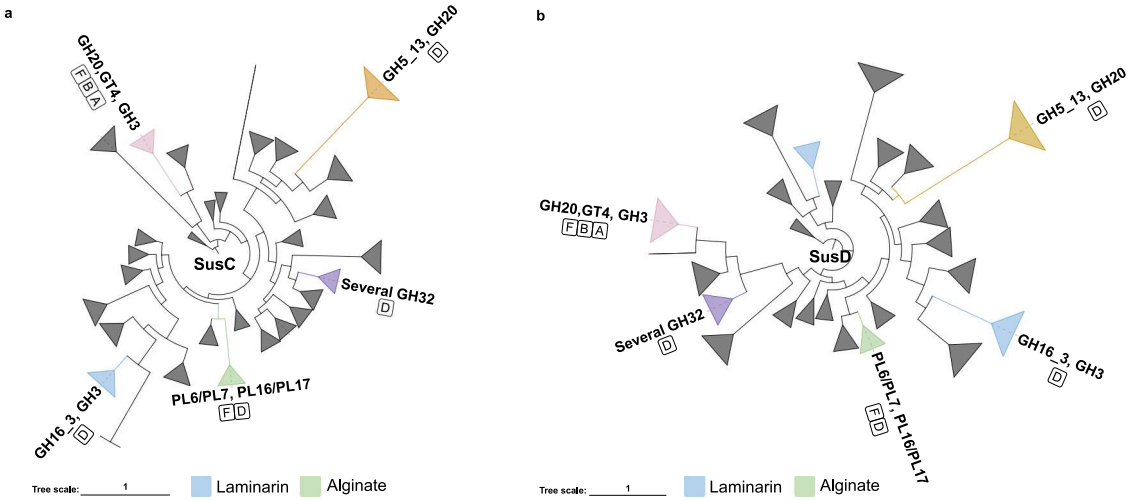
Supplementary Figure S18. Composition of conserved and non-conserved glycoside hydrolase gene family annotations from the three selected genus-representatives. The annotation of gene families was derived from agreements between HMMscan against the dbCAN v9 database and Diamond blastp search against the CAZy database. Genus values presented were derived from the average of the gene counts of the three selected genus-representatives. Conserved genes were required to be present in all three representatives.



Supplementary Figure S19. Composition of transporter genes derived from the three selected genus-representative species. The gene number values shown arethe average of the the three representatives from each genus. The annotation of transporter genes was performed with Prokka v1.14.6 [66].

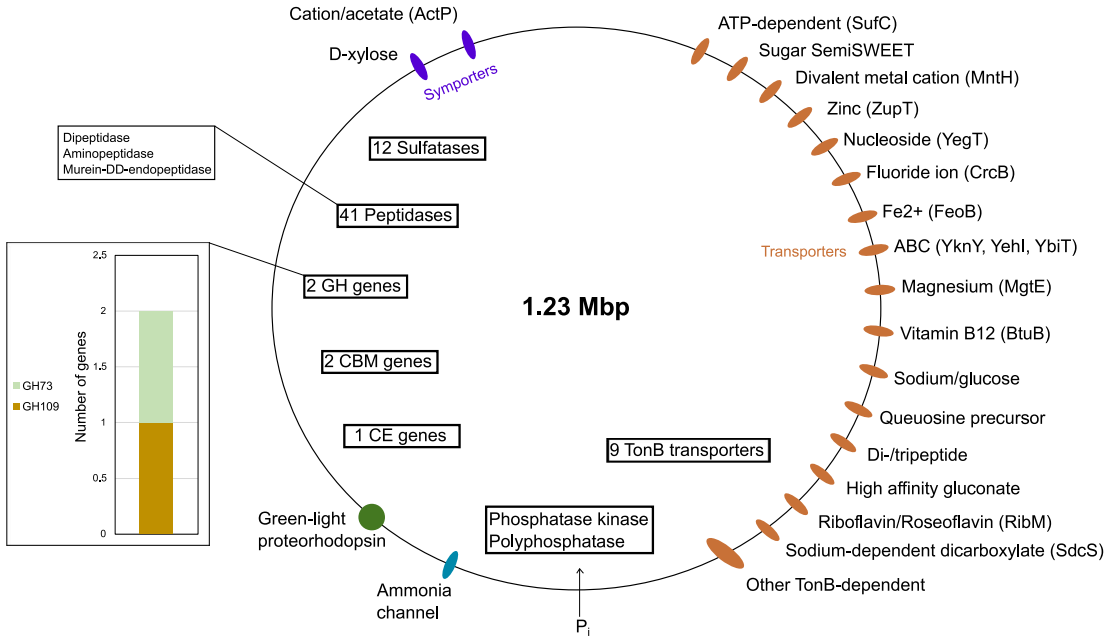


Supplementary Figure S20. Hierarchical clustering analysis of substrate utilisation gene composition of species-representative MAGs. Gene composition was converted to a Bray-Curtis dissimilarity matrix prior to clustering. a) Composition of CAZyme, sulfatase, peptidase and TonB-dependent transporter genes, b) Composition of CAZyme, sulfatase and TonB-dependent transporter genes, c) Composition of CAZyme and

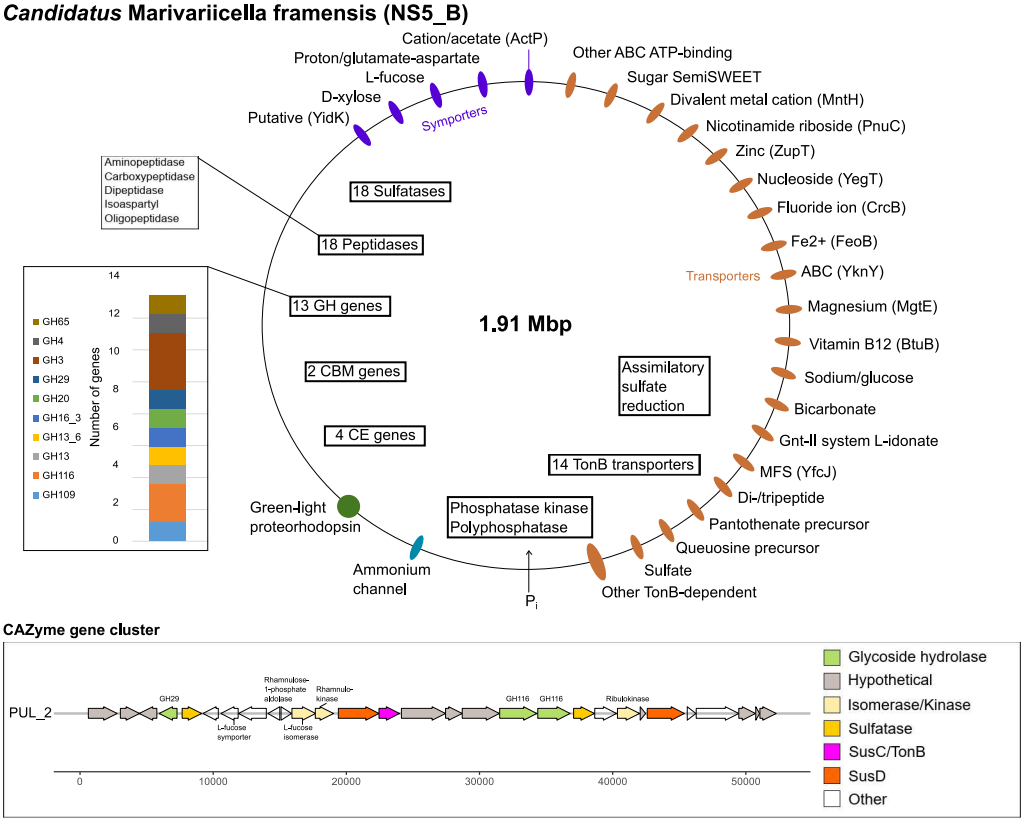


Supplementary Figure S21. Protein trees of SusC/SusD genes identified on polysaccharide utilisation loci of NS5 species representatives and other flavobacteria. a) Susc, b) SusD amino acid sequences from NS5 species along with those from previously published datasets of flavobacteria MAGs and cultured isolates were aligned used MAFFT L-INS-I and trees calculatedusing FastTree. Clusters that are coloured are those that include NS5-derived sequences and conserved gene colocalisations, indicated in the labels. The lettering underneath the cluster labels represent the genus affiliation of the NS5 species within that cluster. Blue branch containing GH16_3 and GH3 represents laminarin-targeting PULs whereas the green branch containing PL6/PL7 and PL16/PL17 represents alginate-targeting PULs.

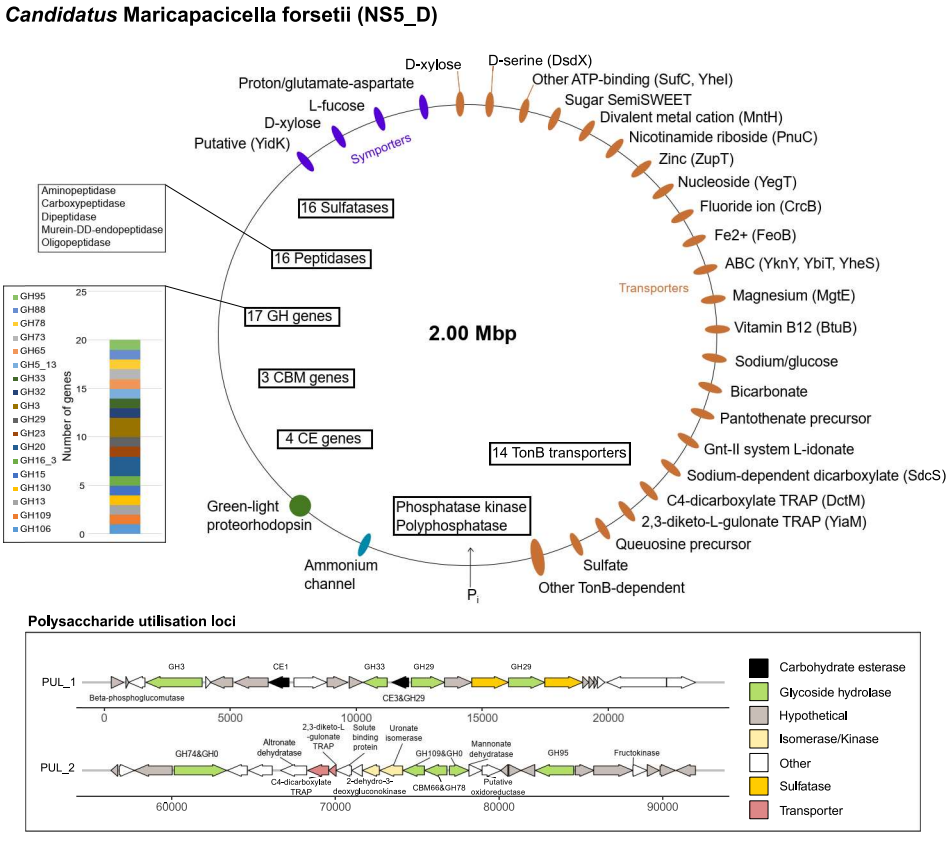
Candidatus Marisimplicoccus framensis (NS5_A)



Supplementary Figure S22. Summary schematic of metabolism of type species *Candidatus* Marisimplicoccus framensis.

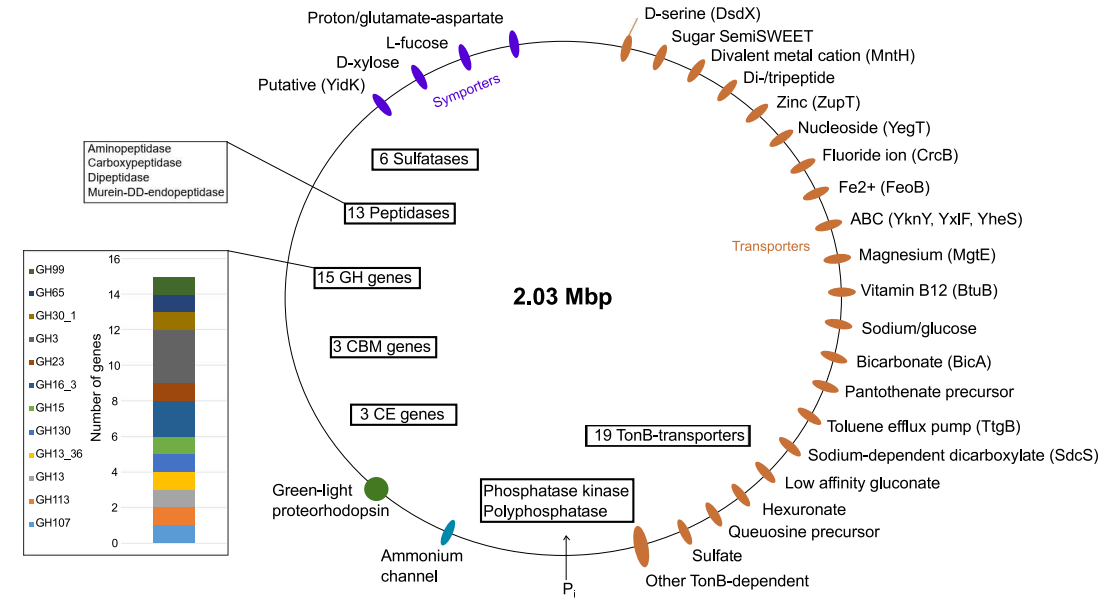


Supplementary Figure S23. Summary schematic of the metabolism of type species *Candidatus* Marivariicella framensis

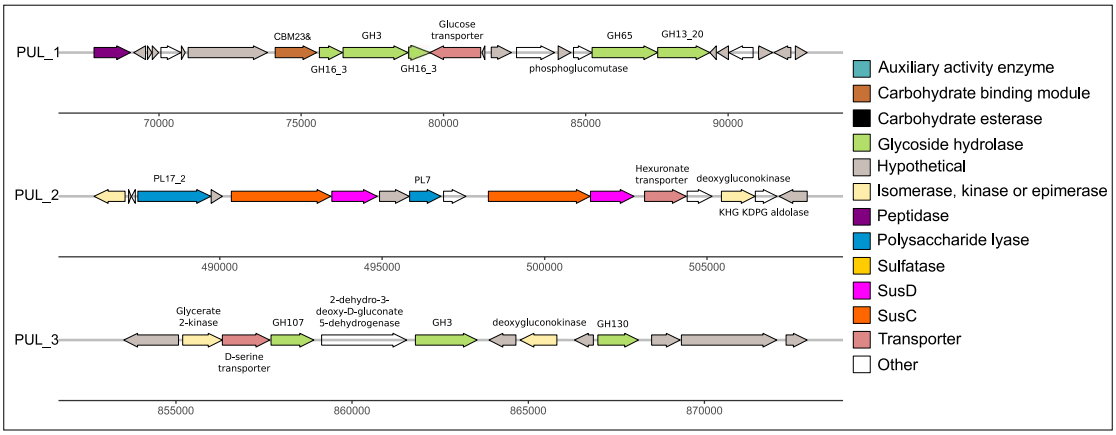


Supplementary Figure S24. Summary schematic of the metabolism of type species *Candidatus* Maricapacicella forsetii

***Candidatus*Arcticimaribacter forsetii (NS5_F)**



Polysaccharide utilisation loci



Supplementary Figure S25. Summary schematic of the metabolism of type species *Candidatus* Arcticimaribacter forsetii

Chapter IV

Variations in Atlantic water influx and sea-ice cover drive taxonomic and functional shifts in Arctic marine bacterial communities

Taylor Priest, Wilken-Jon von Appen, Ellen Oldenburg, Ovidiu Popa, Sinhué Torres-Valdés, Christina Bienhold, Katja Metfies, Bernhard M. Fuchs, Rudolf Amann, Antje Boetius, Matthias Wietz

Manuscript under review in ISME Journal and available on BioRXiv
(https://doi.org/10.1101/2022.08.12.503524)

Contribution of the candidate in % of the total work

Experimental concept and design – 10%

Experimental work/acquisition of experimental data – 5%

Data analysis and interpretation – 60%

Preparation of figures and tables – 90%

Drafting of the manuscript – 70%

Variations in Atlantic water influx and sea-ice cover drive taxonomic and functional shifts in Arctic marine bacterial communities

Taylor Priest[1], Wilken-Jon von Appen[2], Ellen Oldenburg[3], Ovidiu Popa[3], Sinhué Torres-Valdés[2], Christina Bienhold[1,2], Katja Metfies[2], Bernhard M. Fuchs[1], Rudolf Amann[1], Antje Boetius[1,2,4], Matthias Wietz[1,2**]

[1] Max Planck Institute for Marine Microbiology, Bremen, Germany
[2] Alfred Wegener Institute Helmholtz Centre for Polar and Marine Research, Bremerhaven, Germany
[3] University of Düsseldorf, Düsseldorf, Germany
[4] MARUM, University of Bremen, Bremen, Germany

****Corresponding author**:
Matthias Wietz
matthias.wietz@awi.de

ABSTRACT

The Arctic Ocean is experiencing unprecedented changes as a result of climate warming, necessitating detailed analyses on the ecology and dynamics of biological communities to understand current and future ecosystem shifts. Here we show the pronounced impact of Atlantic water influx and sea-ice cover on bacterial communities in the East Greenland Current (Fram Strait) using a multi-year, high-resolution amplicon dataset and an annual cycle of PacBio HiFi read metagenomes. Densely ice-covered polar waters harboured a temporally stable, resident microbiome, with SAR11, *Polaribacter*, SAR92 and SAR86 representatives being the most prominent. In contrast, low-ice cover and Atlantic water influx shifted community dominance to seasonally fluctuating populations that are functionally linked to phytoplankton-derived organic matter degradation, including *Luteolibacter, Flavobacterium* and *Colwellia*. Sea ice cover had the strongest influence on bacterial community functionality. Under high ice cover, communities were enriched in genes for bacterial- and terrestrial-derived organic matter degradation, such as D-amino acids and ketones, and inorganic substrate metabolism. Under low ice cover, genes for the metabolism of phytoplankton-derived carbohydrates, such as laminarin, and dissolved organic nitrogen and sulfur compounds, such as taurine and trimethylamine, were enriched. We subsequently identified populations that are

signatures of distinct environmental conditions and predicted their ecological niches, which included representatives of taxa lacking previous functional descriptions such as the Arctic97B-4 Marine group (*Verrucomicrobiae*). Our study indicates progressive “Biological Atlantification” in the Arctic Ocean, where niche space of Arctic bacterial populations will diminish, while communities that taxonomically and functionally resemble those in temperate oceans will become more widespread.

INTRODUCTION

The Arctic Ocean is experiencing unprecedented changes as a result of climate warming. Of particular significance is the rapid decline in sea-ice extent and thickness [1, 2], with future projections indicating frequent ice-free summers by 2050 [3]. In the Eurasian Arctic, accelerated rates of sea-ice decline have been linked to an increase in heat transport from inflowing Atlantic waters [4], which together weaken water column stratification and increase vertical mixing. The expanding influence of Atlantic water in the Arctic Ocean, termed Atlantification, not only impacts hydrographic conditions but also provides avenues for habitat range expansion of temperate organisms [5, 6]. The consequences of such perturbations on Arctic Ocean ecology are expected to be considerable. In order to predict and understand future changes in the ecosystem state and functioning of the Arctic Ocean, research on the ecology and dynamics of biological communities at the interface between Arctic and temperate oceans is essential [7].

In seasonally ice-covered areas of the Arctic Ocean, bacterial communities exhibit seasonal temporal dynamics, similar to those in temperate ecosystems. These patterns are driven by pulses of organic matter released from phytoplankton blooms [8] and the melting of first year sea-ice [9]. In recent decades, declining sea-ice cover has extended the growing season and increased open-water habitat space of phytoplankton, resulting in a 30% increase in net annual primary production between 1998 and 2012 [10] in shelf and slope areas of the Arctic. Phytoplankton bloom phenology is also shifting, with secondary autumn blooms now being observed in seasonally ice-covered areas [11]. These changes will alter the production and availability of organic matter to bacterial communities over spatial and temporal scales. Recent evidence has shown that sea-ice dynamics also influence the availability of organic matter in surface waters and the transport of carbon and microorganisms to the deep-sea [12–14]. Generally, ice margins are highly productive, because of the combination of early light availability, stratification and diatom-based blooms, which produce large particles and result in relatively high carbon export [15]. However, strong melt events can intensify stratification, resulting in low nutrient supply and delayed export. For example, such a phenomenon in the Fram Strait was found to slow the biological carbon pump by up to 4 months compared to an ice-free situation [12]. Furthermore, declining sea-ice and warming Atlantic waters favour

smaller flagellates, resulting in a pelagic retention system with reduced total annual export [16]. Thus, understanding how sea-ice dynamics influence bacterial communities will provide insights into future biological and biogeochemical changes.

The Fram Strait, the main deep-water gateway between the Arctic and Atlantic Oceans, is a key location for conducting long-term ecological research over environmental gradients and under changing conditions [17]. Fram Strait harbours two major current systems; the East Greenland Current (EGC), which transports polar water southwards in its upper layer, and the West Spitsbergen Current (WSC) that transports Atlantic water northward. The EGC accounts for the export of ~50% of freshwater and ~90% of sea-ice from the central Arctic Ocean and carries Arctic hydrographic signatures [18]. Large-scale recirculation of Atlantic water (AW) into the EGC by eddies is a continuously occurring process, although the magnitude varies across latitudes and over temporal scales [19, 20]. The mixing of these water masses in the marginal ice zone (MIZ) creates different hydrographic regimes reflective of Arctic, mixed water and Atlantic conditions, which can harbour unique bacterial compositions [21, 22]. Carter-Gates *et al.* [21] predicted that future Atlantification of the Arctic may result in a shift towards temperate, Atlantic-type communities. However, further assessments of microbial population dynamics across different temporal and spatial scales, i.e. under Arctic vs Atlantic conditions, are needed to validate such hypotheses.

Here, we performed a high-resolution analysis of the temporal variation of bacterial taxonomy and function at two locations in the EGC between 2016-2018 (MIZ) and 2018-2020 (core EGC), covering the full spectrum of ice cover, daylight and hydrographic conditions (Arctic to Atlantic water masses). Our study is embedded in the “Frontiers in Arctic Marine Monitoring” (FRAM) Ocean Observing System that employs mooring-attached sensors and autonomous Remote Access Samplers (RAS) to continuously monitor physicochemical parameters and biological communities in the Fram Strait. This analysis encompasses a 2 two-year 16S rRNA amplicon dataset supplemented with an annual cycle of PacBio HiFi read metagenomes, expanding a previous assessment of microbial dynamics over a single annual cycle in the EGC [23]. We hypothesise that high AW influx and low sea-ice cover will result in bacterial communities dominated by chemoheterotrophic populations, resembling those of temperate ecosystems. Our investigation provides essential insights into the effects of the changing Arctic on marine microbial ecology and biogeochemical cycles.

RESULTS AND DISCUSSION

The amplicon dataset incorporates samples (>0.2 µm fraction) collected at weekly to biweekly intervals at the Marginal Ice Zone (MIZ; 2016 - 2018) and in the central EGC (core-EGC; 2018 - 2020), between 70 – 90 m depth (Supplementary Table S1). The two locations were selected in order to capture the full spectrum of water mass and sea ice conditions. The core-EGC

featured almost year-round dense ice cover (in the following abbreviated with “high-ice”) and polar water (PW) conditions. In contrast, the MIZ featured variable, generally lower ice cover (in the following abbreviated with “low-ice”) and periodic Atlantic water (AW) influx (Figure 1). To provide a more visual representation, animated GIFs were created for current velocities at the depth of sampling (Supplementary Figure S1) and sea-ice cover dynamics (Supplementary Figure S2) over the four year period. By combining the high-resolution data from both mooring locations, we are able to assess bacterial community dynamics over temporal scales and in relation to Arctic- and Atlantic-dominated conditions.

Bacterial community and population dynamics over temporal scales

Combining the two, 2-year amplicon datasets resulted in 3988 non-singleton Amplicon Sequence Variants (ASVs) (Supplementary Table S2) being recovered, which were initially used in a taxonomy-independent approach to assess community dynamics over environmental gradients. Bacterial community composition shifted with environmental conditions (Figure 2). A stepwise significance test identified AW proportion, daylight and past ice cover (average ice cover of the days preceding the sampling event) as the significant factors explaining compositional variation (model $R^2 = 0.23$, p-value = 0.001) (Supplementary Table S1). AW proportion explained 13% of the total variation, compared to 6% for daylight and 4% for past ice cover. The pronounced impact of AW reflects previous observations that different water masses harbour distinct bacterial assemblages [21, 25, 26] with important implications for the future Arctic Ocean, as AW influence is expected to expand. The impact of daylight can be directly and indirectly linked to bacterial community dynamics through their own phototrophic capacities and via phytoplankton dynamics. In our study, changes in light availability reflect seasonal dynamics, which are well evidenced in temperate and polar ecosystems [27–29] and also occur in ice-free areas of the Fram Strait [23]. The impact of ice cover on the bacterial community is in line with previously reported ecosystem responses and reflects the influence that it has on hydrographic and physiochemical conditions as well as organic matter availability [9, 12].

To gain a more detailed understanding on bacterial community structuring, the dynamics of ASVs across samples was assessed. In total, 75% of the ASVs were detected at both mooring sites, whilst 16% and 8% were unique to the MIZ and core-EGC, respectively. The frequency of detection and maximum relative abundance of ASVs shared by both sites exhibited a strong positive linear relationship, i.e. those identified in more samples also reached higher maximum relative abundances (Figure 3a). To facilitate further comparisons, we categorised ASVs into three groups: a) resident ASVs (Res-ASVs), present in >90% of samples, b) intermittent ASVs (Int-

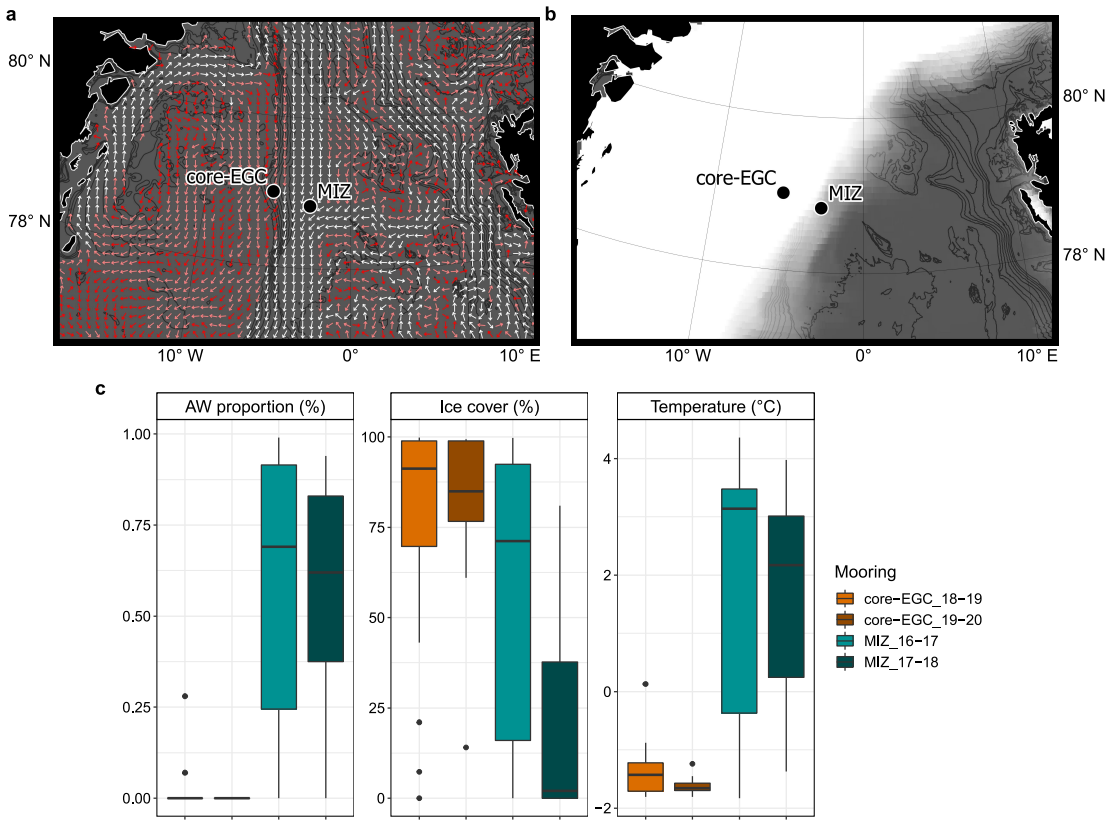


Figure 1. Geographical location of seafloor moorings and variation in environmental conditions in MIZ (2016–2018) and core-EGC (2018–2020). a) Example representation of monthly average (January 2020) current velocities at the approximate depth of water sampling (78m). White and dark red arrows indicate strongest and weakest velocities, respectively. b) Example representation (December 2019) of sea-ice cover. Increasing opacity of white colour reflects increasing sea-ice cover, where pure white = 100% sea-ice cover. Values for current velocities and sea-ice concentration were obtained from copernicus.eu under the 'ARCTIC_ANALYSIS_FORECAST_PHY_002_001_a'. c) Boxplots illustrating variation in AW proportion, ice cover and temperature at the moorings. The bathymetric map was made using publically available bathymetry data from GEBCO [17].

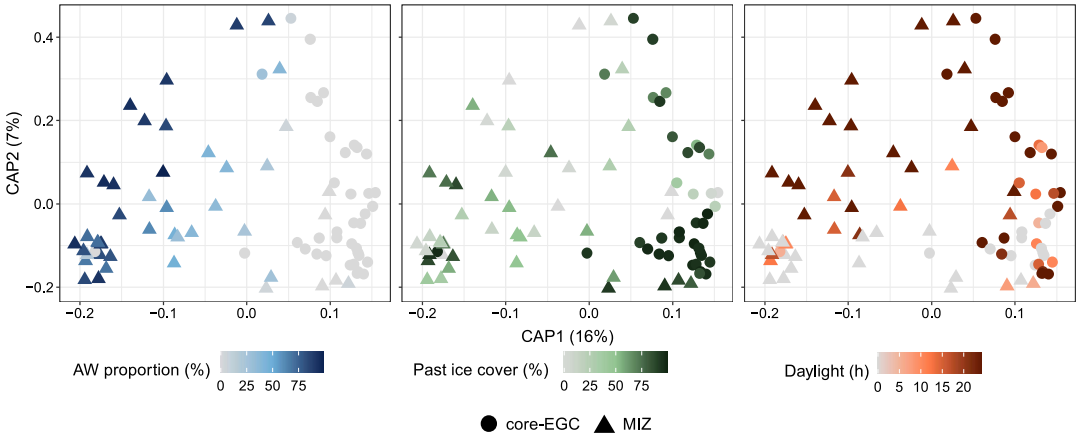


Figure 2. Community structure across variations in water mass, ice cover and daylight conditions. Distance-based redundancy analysis based on Bray-Curtis dissimilarities of community composition along with AW proportion (blue symbols), past ice cover (green symbols) and daylight (orange symbols) as constraining factors. The factors were selected using a stepwise significance test and combined into a single model (R2 = 0.1, pvalue = 0.01) that constrains 14% of the total variation. For ease of interpretation, the environmental conditions are visualised individually on the same ordination.

ASVs), present in 25 - 90% of samples, and c) transient ASVs (Trans-ASVs), present in <25% of samples. There were only 232 Res-ASVs, but these represented the largest proportion of the sampled bacterial communities (48 - 88% relative abundance). In comparison, the 1904 Int-ASVs constituted 12 – 48% and the 1852 Tran-ASVs between 0.4 – 5.8% of relative abundances. Presence of a dominant resident microbiome, represented by a minority of ASVs, corresponds to previous observations in summertime Fram Strait samples [30] as well as the Western English Channel and Hawaiian Ocean time-series [27, 31].

The fluctuations in abundance of the three community fractions were primarily associated with changes in AW proportion, with strong negative correlations for the resident (Pearson's coefficient: -0.50, p-value <0.01) and transient fractions (Pearson's coefficient: -0.32, p-value < 0.01) and strong positive correlations for the intermittent fraction (Pearson's coefficient 0.57, p-value <0.01). This is evident in the more stable temporal dynamics for the resident community in core-EGC compared to MIZ samples (Figure 3c). The resident microbiome was phylogenetically diverse, incorporating both abundant and rare community members. Res-ASVs were assigned to 47 families and 79 genera, with the *Flavobacteriaceae* (n=15)*, Magnetospiraceae* (n=13)*, Marinimicrobia* (n=11) and SAR11 Clade I (n=22) and Clade II (n=17) harbouring the largest diversity. Maximum relative abundances of Res-ASVs ranged from 0.035 – 13.9%, with the most prominent being affiliated with SAR11 Clade Ia (asv1 - 14%), *Polaribacter* (asv6 - 14%), *Aurantivirga* (asv7 - 12%), SUP05 (asv2 - 12%), SAR92 (asv16 - 11%) and SAR86 (asv3 – 9%). Pronounced fluctuations of the intermittent community (11 – 48% relative abundance) coincided with AW influx at the MIZ. The intermittent community was more phylogenetically diverse than the resident microbiome, encompassing 250 genera, and also comprised rare and abundant populations that reached between 0.004 – 15% maximum relative abundance. The most diverse taxa included the SAR11 Clade II (n=148), *Marinimicrobia* (n=129), NS9 Marine Group (n=78), AEGEAN-169 (n=73) and *Nitrospinaceae* (n=56). Those reaching the largest relative abundances were affiliated with *Luteolibacter* (asv24 - 15%), *Flavobacterium* (asv140 - 10%), *Polaribacter* (asv206 - 10%) and *Colwellia* (asv89 - 9%). The resident and intermittent community fractions shared 71 genera, which constitutes 90% of the genus-level diversity of the resident microbiome. Hence compositional changes over temporal scales are driven by dynamics on the species- and population-level.

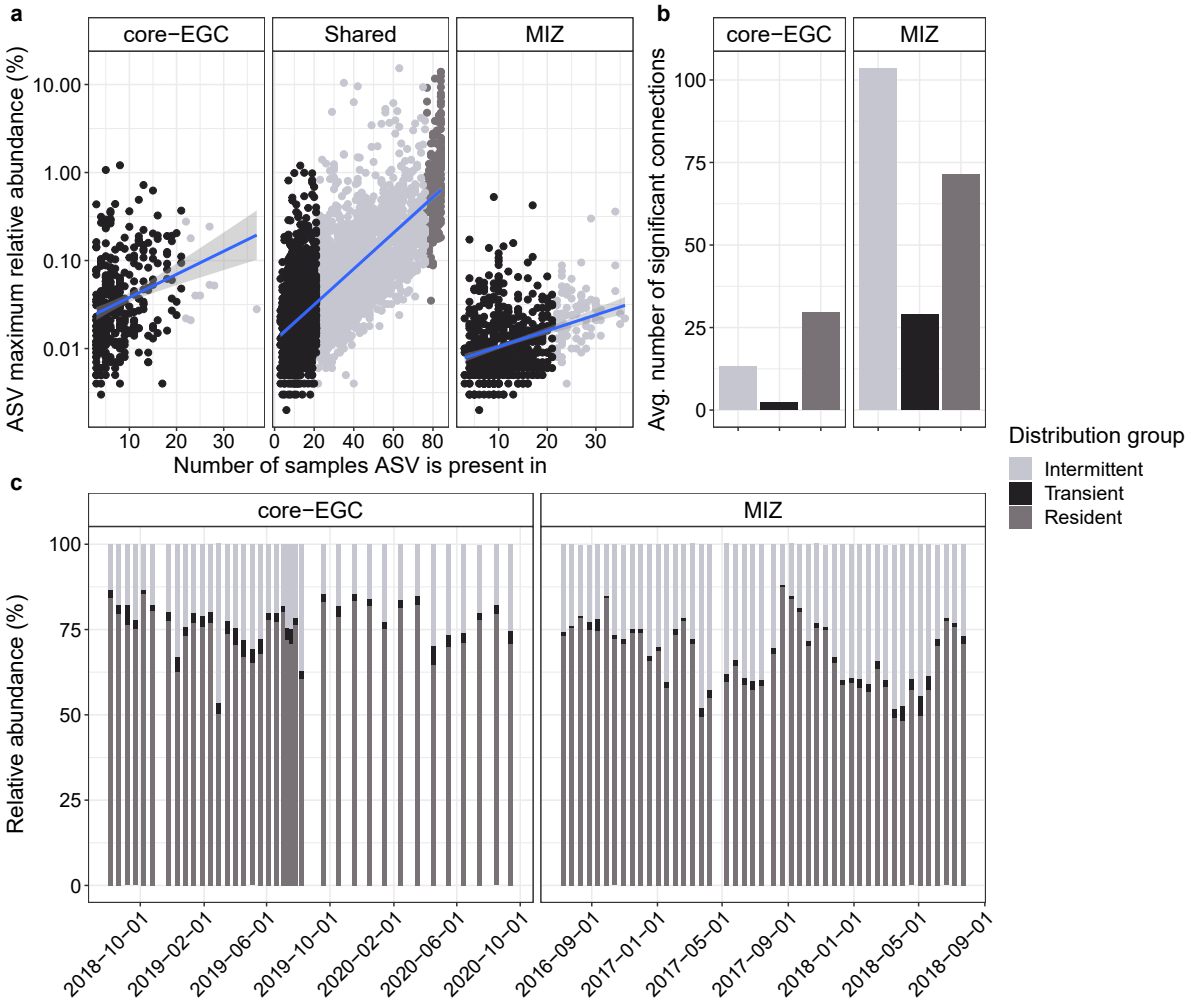


Figure 3. Distribution dynamics and co-occurence of ASVs. a) The occurence of ASVs across samples in relation to their maximum relative abundances along with categorization into resident, intermittent and transient. b) Average number of connections within the co-occurrence networks for resident, intermittent and transients ASVs. c) Relative abundance dynamics of the resident, intermittent and transient ASVs over time.

To consolidate the observed community structuring and further illuminate ASV dynamics, we computed co-occurrence networks and contextualised them with environmental conditions (Supplementary Figure S3). The networks further supported a greater microbiome stability in the central EGC, with the core-EGC network being enriched in co-occurring Res-ASVs whilst more Int-ASVs had significant co-occurrences in the MIZ network (see Supplementary Information S1 for a more detailed description).

Overall, the seasonally and spatially variable conditions in the MIZ drive substantial dynamics of distinct bacterial populations. Accordingly, the environmentally less dynamic core-EGC is reflected in a more stable resident microbiome that fluctuates less with seasons, likely reflecting adaptations to polar water and nearly year-round ice cover.

Taxonomic signatures of distinct environmental conditions

A sparse partial least squares regression analysis (sPLS) identified 430 ASVs that were associated with distinct environmental conditions. These ASVs formed eight distinct clusters based on similar, significant correlations to seven environmental parameters (Figure 4a). The composition of ASVs in each cluster were largely unique, revealing distinct taxonomic signatures associated with specific environmental conditions (Figure 4b). The three largest clusters incorporated 88% of the ASVs and were separated based on their associations to different water mass and ice cover conditions. Clusters C1 and C2 represented AW conditions, with C1 also being associated with low-ice cover. In contrast, cluster C8 represented polar water (PW) conditions under high-ice cover. In accordance with the distribution dynamics described above, the AW-associated clusters comprised a higher proportion of Int-ASVs, 51 – 88%, compared to ~50% Res-ASVs in PW-associated clusters. An additional five smaller clusters (C3 - C7) were also identified that corresponded to polar day and polar night under different ice cover and water mass conditions. Comparing the most prominent ASVs of each

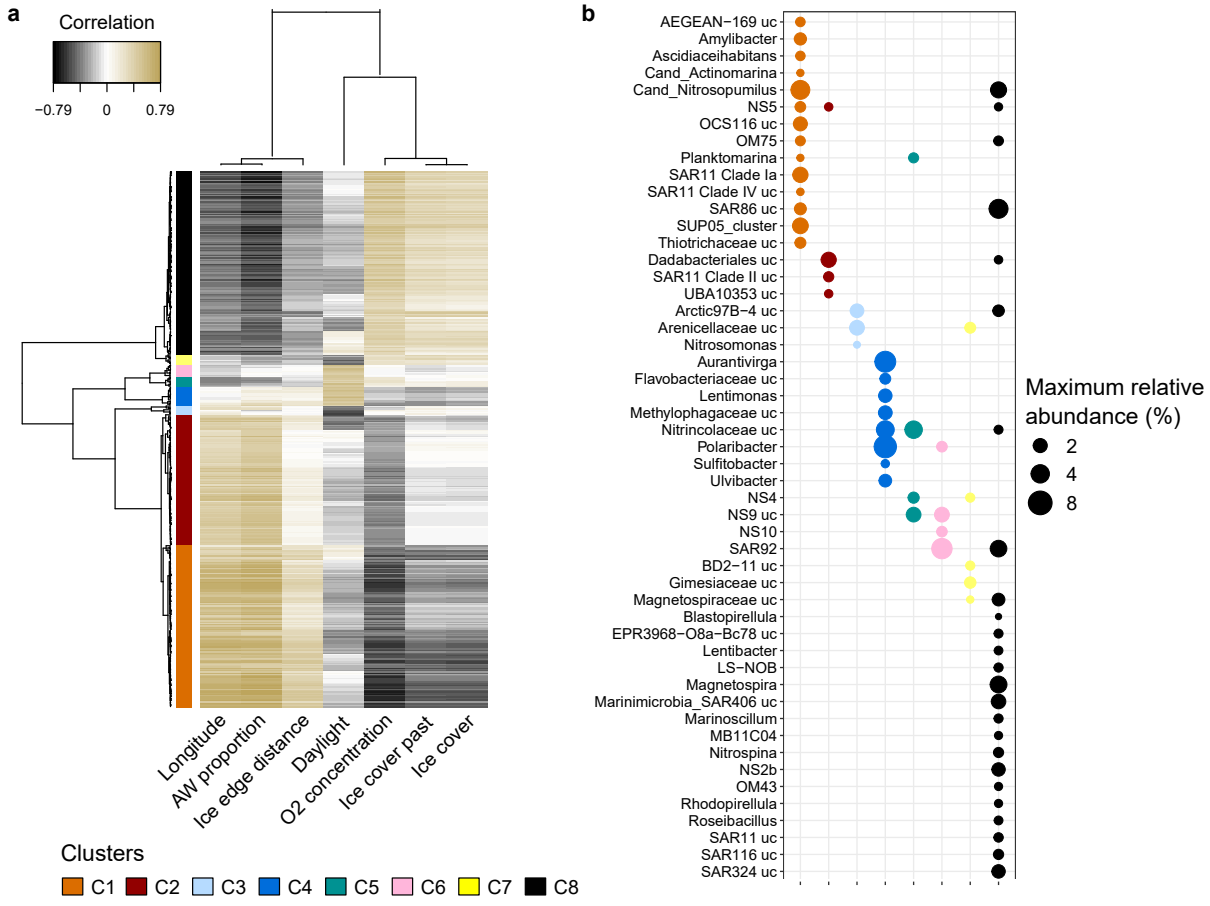


Figure 4. **Sparse partial least square regression linking community structure and environmental parameters.** a) Heatmap showing the eight major sPLS clusters identified that encompass 430 ASVs with significant correlations to environmental conditions. b) Representation of the most prominent genera in each cluster. ASVs that did not reach >1% were excluded from the taxonomic comparison, whilst the remaining were grouped by genus and the maximum abundance of each genus shown. Due to high collinearity with AW proportion, temperature and salinity were excluded from sPLS analysis. Thresholds: coefficients >0.4, p-values <0.05.

cluster (reaching >1% relative abundance) revealed unique taxonomic signatures at the genus-level (Figure 4b). For instance, *Amylibacter*, SUP05 Clade and AEGEAN-169 were signatures of the AW-associated, low-ice cluster C1, whilst the SAR324 Clade, NS2b Marine Group and *Magnetospira* were signatures of the PW-associated, high-ice cluster C8. Overall, this pattern underlines water mass and ice cover as major drivers of community structure, whilst a smaller number of ASVs are strongly influenced by daylight and seasonality.

Signature populations associated with distinct environmental conditions

We contextualized ASV dynamics with metagenome-assembled genomes (MAGs) to link distribution with metabolic potential and predict ecological niches of populations. From nine PacBio HiFi read metagenomes, derived from the 2016-17 annual cycle in the MIZ, we recovered 43 manually-refined, population-representative MAGs (delineating cut-off of 99% average nucleotide identity; ANI). The MAGs represented 26 – 49% of the metagenomic reads. Of these, twelve were high-quality drafts whilst the remainder were medium-quality drafts [32], but all MAGs were highly contiguous (average number of contigs = 33) and >80% contained at least one complete rRNA gene operon. The MAGs represented a broad phylogenetic diversity, including 35 genera, 27 families and nine classes (Supplementary Figure S4 and Supplementary Table S3). Comparing the MAGs to those recently recovered from the Fram Strait [22] indicated that 32 were novel species (<95% ANI), whilst eleven represented previously recovered populations (>99% ANI).

Through competitive read recruitment, 27 of the MAGs were linked to distinct ASVs (based on 100% identity threshold). Of these, 18 could be associated with sPLS clusters and thus distinct environmental conditions. These 18 representatives of sPLS clusters, which encompass an ASV and MAG, are hereon referred to as “signature populations” (Table 1 and Figure 5). Signature populations included some of the most prominent representatives of each cluster, such as asv6-*Polaribacter* and asv7-*Aurantivirga* from cluster C4 and asv18-SAR86 Clade from cluster C8. Based on their dynamics being driven primarily by ice cover and water mass, we define signature populations of cluster C1 and C2 as Atlantic signatures and those of C8 as Arctic signatures.

For consistency, signature populations will be identified by their asv number, with the corresponding MAG information provided in Table 1.

To corroborate the associations between signature populations and distinct environmental conditions, we assessed their distribution across the Tara Oceans Arctic dataset. Signature populations of polar day clusters were more prominent in the euphotic zone, whilst Arctic and polar night signature populations increased below 100 m depth (Supplementary Figure S5). Furthermore, signature populations identified as residents were present, on average, higher relative abundances than intermittent or transient populations.

These patterns are in agreement with our observations from the EGC and reflect the sampling campaign of Tara Oceans in the Arctic, which focused on above continental shelf locations during summer (more detailed comparisons are provided in Supplementary Information S1).

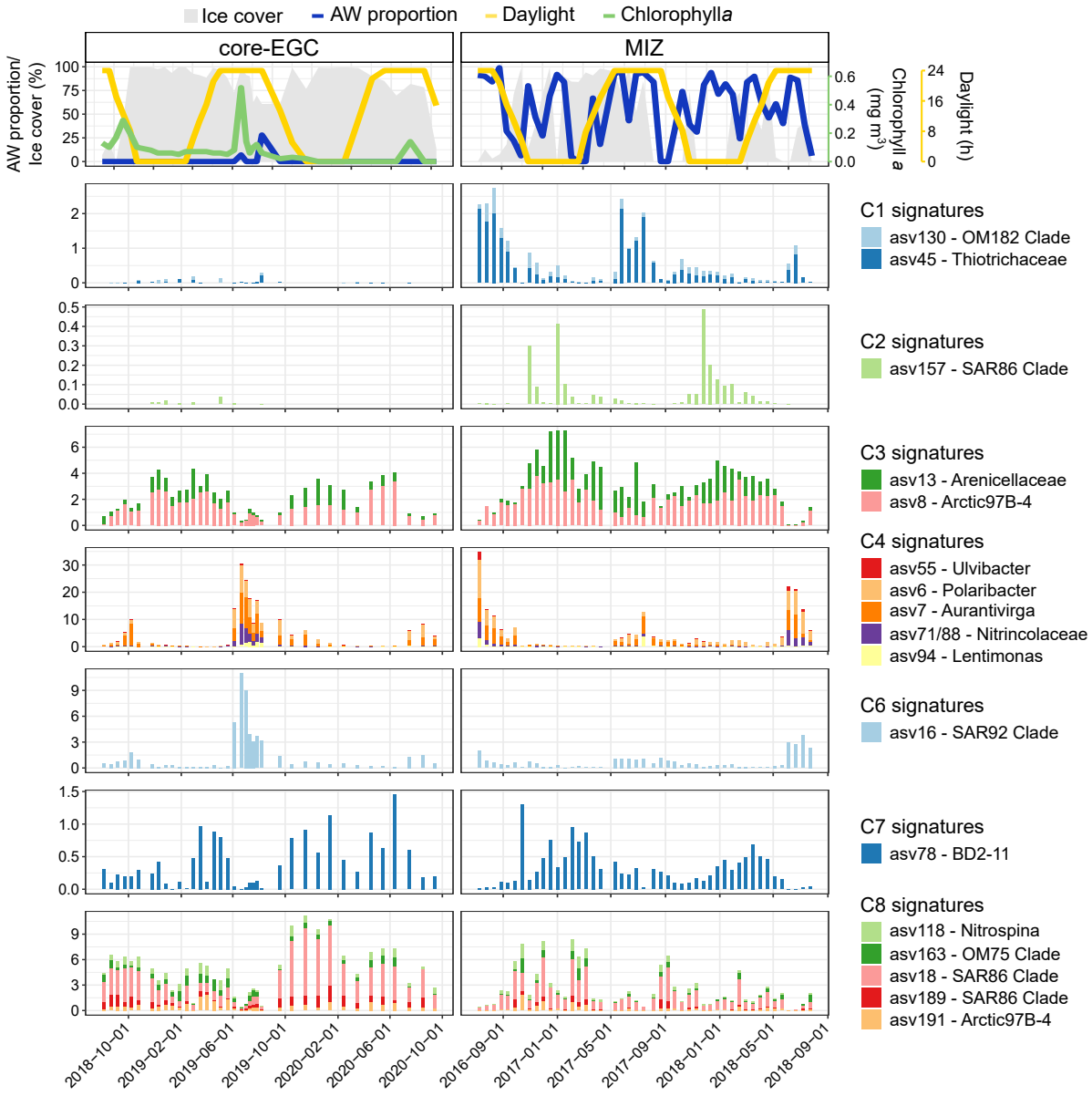


Figure 5. Temporal dynamics of signature populations. Signature populations were identified as ASV representatives of sPLS clusters that a correpsonding MAG was recovered for (based on 100% identity threshold competitive read recruitment). The temporal dynamics visualized are derived from ASV data only. The missing chlorophyll a data in 2016-18 is due to the lack of a sensor on the MIZ mooring.

Ecological niches of signature populations

By assessing temporal dynamics in functional genes, we are able to predict the ecological niches of signature populations within the context of environmental conditions. Of particular interest were the signature populations of Atlantic (clusters C1 and C2) and Arctic (cluster C8) conditions, as their ecology provides insights into how bacterial community structure may shift

in the future Arctic Ocean. Ecological descriptions of daylight-driven clusters and tables with complete functional annotations for all signature populations are provided in Supplementary Information S1 and Supplementary Files S1, respectively.

Atlantic signature populations

Atlantic signature populations were affiliated with the *Gammaproteobacteria* class, and included representatives of OM182 (UBA4582), *Thiotrichaceae* (GCA-2705445) and SAR86 (SCGC-AAA076-P13). All three populations reached significantly higher relative abundance values in MIZ compared to core-EGC, however, within the MIZ, distinct dynamics were observed (Figure 5). The *Thiotrichaceae* (asv45) population peaked during polar day whilst the SAR86 (asv157) population peaked during polar night. The dynamics of the OM182 (asv130) population was related only to AW and not to daylight. Functional gene predictions revealed unique lifestyles and metabolic capacities for each population. The following analyses focuses on the asv45 and asv130 populations, with more details about the asv157 population provided in Supplementary Information S1.

The *Thiotrichaceae* population harbours the capacity to utilise phytoplankton-derived organic compounds, the availability of which likely stimulates growth under polar day conditions. These include methanethiol and C1 compounds. Methanethiol originates from the demethylation of dimethylsulfoniopropionate (DMSP) [33], an osmoprotectant produced by phytoplankton. DMSP concentrations in the Arctic Ocean show spatial variation and are influenced by water mass and sea-ice, with highest concentrations reported in areas directly influenced by Atlantic water inflow (western Eurasian Arctic) [34]. The concentration of DMSP in these regions is tightly coupled to chlorophyll *a* [34, 35]. As such, methanethiol is likely more available in Atlantic waters under polar day conditions. The oxidation of methanethiol, through the methanethiol oxidase gene (MTO), results in the production of formaldehyde and hydrogen sulfide. In the asv45 population, we identified the MTO gene along with a complete tetrahydromethanopterin (H4-MPT)-dependent oxidation pathway and sulfide oxidation machinery (dsrAB and soeABC). Combined, this genetic repertoire would allow the asv45 population to use methanethiol as a carbon, sulfur and energy source.

A similar metabolism has been reported for members of the *Rhodobacteraceae,* which are capable of degrading methanethiol and subsequently oxidising the hydrogen sulfide for energy generation [36]. The H4-MPT-dependent formaldehyde oxidation pathway has been traditionally affiliated with methanogenic bacteria but was also demonstrated in methylo- and methanotrophic members of the Alpha- and *Gammaproteobacteria* [37]. To our knowledge, this is the first such description in a sulfur-oxidizing member of the *Thiotrichaceae* family. Although experimental evidence is needed to consolidate these findings, we confirmed the presence of the above-described pathways in each of the species-representative genomes

from the assigned GTDB genus (GCA-2705445). The GCA-2705445 genus contains several representatives that are classified as *Thiothrix* in NCBI. The distinct metabolic features described above may represent unique characteristics and, in line with the GTDB classification, suggests that GCA-2705445 species are distinct from other *Thiothrix*.

The OM182 population differed from the other AW-associated signatures by showing daylight-independent dynamics. Functional gene annotations indicated a motile lifestyle with the capacity to oxidise sulfur and carbon monoxide (CO) as well as degrade taurine and methylamine, thus representing an aerobic, sulfur-oxidising methylotroph. Furthermore, the presence of the complete *sox* system along with polysulfide reductase (*pshAB*) and flavocytochrome c-sulfide dehydrogenase (*fccAB*) genes indicates the capacity to store and use elemental sulfur. The diverse metabolic capacities of the asv130 population may explain the observed dynamics over the time-series, with a capacity to switch nutrient and energy sources under different conditions. For example, under high daylight conditions, CO oxidation combined with the utilisation of organic compounds presumably provides sufficient energy and nutrients for growth. CO production in the oceans is linked to the photolysis of coloured dissolved organic matter and direct production by phytoplankton [38, 39], and thus concentrations would be elevated during polar day and in periods of high productivity. Under such conditions, the capacity to use taurine and methylamine, which are compounds related to phytoplankton production and organic matter degradation respectively, would provide further access to carbon, nitrogen and sulfur. CO oxidation as a supplemental energy source has been previously evidenced in some marine *Gammaproteobacteria* [40], however the dominant organisms performing such processes are typically affiliated with *Rhodobacteraceae* members. In general, only a few heterotrophic populations inhabiting the upper water column have been linked to sulfur- and CO-oxidation. The OM182 Clade may be an important contributor to the biogeochemical cycling of some carbon and sulfur species in the Arctic pelagic environment.

Arctic signature populations

The Arctic signature populations each exhibited highly similar dynamics, with peak relative abundances under high-ice cover, low AW proportion and low daylight conditions (Figure 5). The most prominent of these were the asv18 (SAR86), asv118 (*Nitrospina*) and asv163 (OM75). All five populations harboured distinct metabolic capacities that were either indicative of chemoautotrophic lifestyles or chemoheterotrophic lifestyles, with a capacity to use diverse substrates for growth beyond phytoplankton-derived organic compounds. In this regard, their genetic tools were notably different to Atlantic signature populations. We focus here on SAR86 (due to the high relative abundance) and Arctic97B-4 (due to limited ecological information currently available), with detailed descriptions of other C8 signature populations provided in

Supplementary Information S1. In addition, we assessed the cluster C7 signature population, asv78 – BD2-11, a resident but poorly characterized taxon linked to PW and low-daylight conditions.

The most prominent Arctic signature population, asv18, likely represents a novel genus in the SAR86 Clade, based on <60% average amino acid identity (AAI) to other GTDB representatives. However, it shares >99% AAI to a recently recovered MAG from the Fram Strait (FRAM18_bin252) [22], which corroborates its assignment to the resident microbiome. SAR86 Clade members are known as photoheterotrophs, with distinct ecotypes relating to phototrophic and carbohydrate degradation capacities [41]. They are also known as one of the most prominent gammaproteobacterial responders to spring phytoplankton blooms in temperate ecosystems, wherein they are predicted to use carbohydrates and DMSP [42]. In agreement, our Arctic MAG encodes a green-light proteorhodopsin, typical of lower photic zone-inhabiting organisms [43], as well as carbohydrate degradation genes. However, the large peptidase (n=19) to CAZyme (n=7) gene count ratio and the affiliation of CAZyme gene families with peptidoglycan recycling (GH103 and GH84), suggests a preference for proteinaceous substrates. Furthermore, the population encoded the capacity to metabolise D-amino acids, via conversion to α-keto acids [44]. In conjunction with the phylogenetic distance, these metabolic distinctions to other SAR86 Clade members may represent features of a novel, Arctic-specific genus.

The asv191 signature Arctic population was assigned to the Arctic97B-4 group (*Verrucomicrobiae*), for which no functional information is available to date. 16S rRNA gene-based studies have indicated elevated proportions of Arctic97B-4 in subsurface waters and a tight coupling with other deep water clades, such as SAR202 [45, 46]. An enrichment of Arctic97B-4-affiliated sequences was also identified in the small particle-attached fraction of Southern Ocean samples [47]. In accordance with these findings, the functional annotations of asv191 population suggest a motile chemomixotroph with the capacity to oxidise methane, fix carbon and degrade sulfated carbohydrates. Comparable to other marine *Verrucomicrobia*, the asv191 population encoded a high number of CAZymes (23 genes) and sulfatases (84 genes). However, the peaks of asv191 under no- to low-daylight conditions suggest that alternative substrates are also used. We identified the key marker genes for carbon fixation through the reductive TCA cycle (*korAB*, *por*, *trfAB*) and the aerobic oxidation of methane through formate and formaldehyde (*mdh*, *metF* and *folD*). However, the key methane monooxygenase gene (*pmo*) was not detected. This genetic repertoire is identical to a recently described MAG from the same family (*Pedosphaeraceae*), recovered from a bioreactor community [48]. The authors of that study suggested a potentially novel methane monooxygenase gene that was originally annotated as hypothetical. The aerobic oxidation of methane in the water column is typically associated with environments above continental

shelves and oxygen minimum zones, where methane is supplied from the sediment or anaerobic processes below. Studies from above continental shelves in the Arctic have shown supersaturation of methane, with significantly elevated concentrations under sea-ice compared to ice-free conditions [49, 50]. The increased prevalence of the Arctic97B-4 population under high-ice cover may thus be related to increased methane availability in conjunction with their particle-attached lifestyle, as methane production is known to occur in marine particulate organic matter[51].

The cluster C7 representative, asv78, was assigned to two distinct classes, the BD2-11 terrestrial group (SILVA) and the *Gemmatimonadetes* (GTDB). This discrepancy reflects the relatively recent assignment and largely unresolved phylogeny of the *Gemmatimonadota* phylum. Based on the few available cultured representatives and 16S rRNA-gene based studies, the *Gemmatimonadetes* harbour aerobic/semi-aerobic chemoorganoheterotrophs inhabiting soil environments, but are also reported from freshwater habitats and deep-sea sediments [52, 53]. Their presence in the upper marine water column however, is rarely reported. The asv78 population encodes the capacity to use a wide range of organic substrates for growth, and also perform aerobic denitrification. However, the presence of periplasmic nitrate reductase (*nap*) and nitrite reductase (*nirK*) genes but absence of downstream genes required for further reduction to N_2 suggests an incomplete denitrification pathway. In addition, we identified genes to metabolise taurine, hypotaurine, D-amino acids, dicarboxylic acids and halogenated haloaliphatic compounds. The sources of these compounds in the marine environment vary, with taurine being attributed to phytoplankton and metazoa, D-amino acids to bacteria, and halogenated compounds to all forms of biota. With asv78's capacity to reduce nitrate, this would provide metabolic flexibility to prevail under low daylight conditions and high-ice cover.

Whole community functional shifts with contrasting conditions

To determine if the whole community functionality shows comparable shifts with AW and sea ice dynamics, we assessed differences in functional gene composition on the raw HiFi metagenomic reads. From the nine metagenomes generated here, 17.6 million open reading frames were identified (Supplementary Table S4), of which 54% were assigned a function and 92% were assigned a taxonomy. Expectedly, taxonomic classifications of individual genes, varied in resolution, with 92% being assigned to a kingdom and 37% to a genus. The robustness of classifications was consolidated by comparing community composition recovered from reads to that from the ASVs, which showed high congruence at the class level (Supplementary Figure S6).

A dissimilarity analysis of whole-community functionality separated samples into two distinct clusters, with ice cover being the only statistically significant factor between the two

(F-statistic = 12.6, p-value=0.009) (Figure 6). A total of 1088 differentially abundant genes were identified between the two clusters, with 328 and 845 genes enriched under high- and low-ice conditions, respectively. Enriched genes were related to substrate uptake and degradation (Supplementary Figure S7). Bacterial communities under low-ice exhibited an enhanced capacity to utilize carbohydrates, dissolved organic nitrogen (DON) and sulfur (DOS) compounds, indicative of phytoplankton-derived organic matter. In contrast, high-ice conditions were enriched in genes involved in the metabolism of amino acids, proteins, aromatics and ketone compounds. These patterns reflect fundamental differences in community functionality, corroborating MAG-derived evidence that low-ice communities likely rely on labile organic compounds derived from phytoplankton.

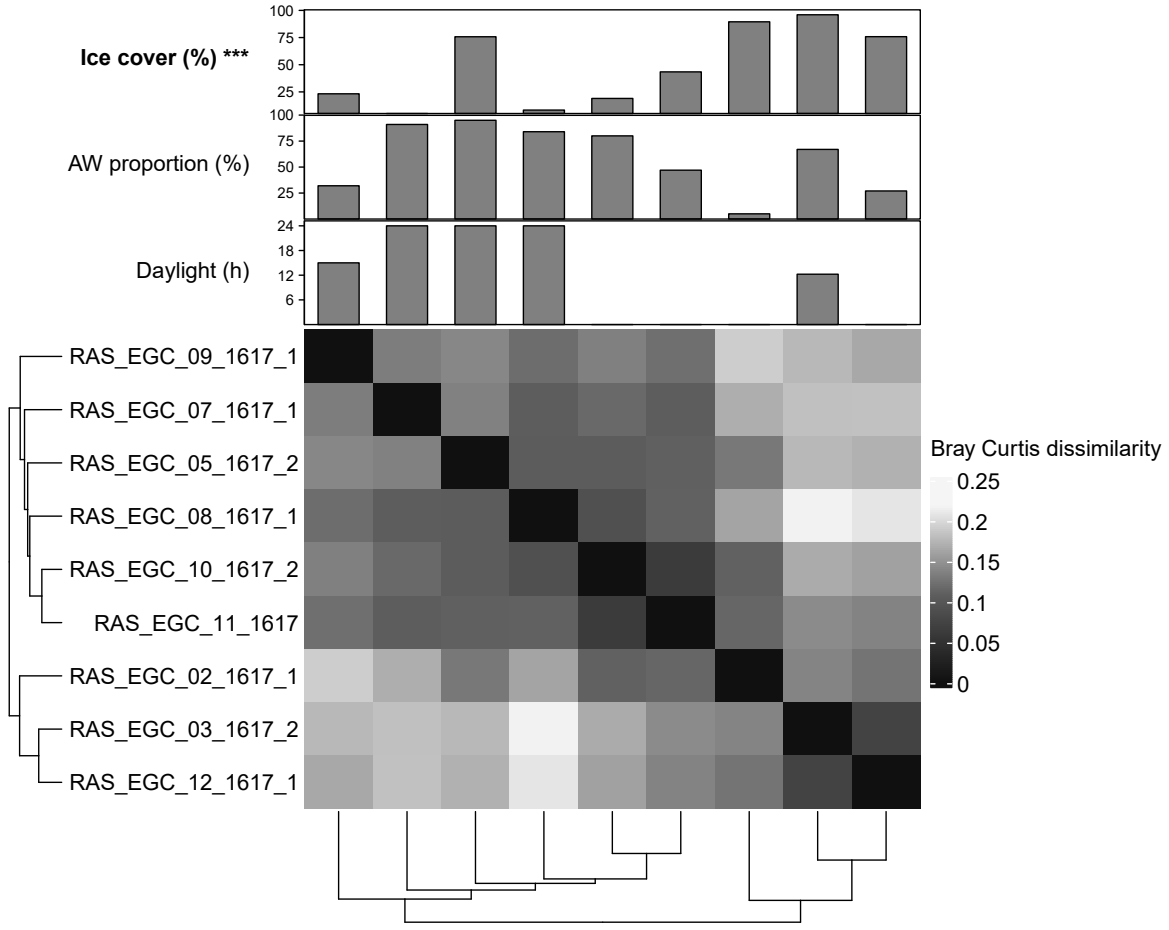


Figure 6. Clustering of sample functional gene composition based on Bray-Curtis dissimilarities.

Functions enriched under low-ice cover

Enrichment of glycoside hydrolase families GH16_3, GH36, GH42 and GH8 indicate an increased potential to degrade laminarin and α-galactose- and β-galactose-containing polysaccharides (Figure 7). In addition, numerous transporters and degradation genes related to mono- and disaccharides were enriched, including for D-xylose, glucose and rhamnose.

Carbohydrates represent a major component of dissolved (15 - 50% [54, 55]) and particulate(3 - 18% [56]) organic matter, and are key substrates for heterotrophic marine bacteria. The major source of carbohydrates in the oceans is phytoplankton, wherein they can constitute up to 50% of the cell biomass[57]. The release of carbohydrates during phytoplankton blooms stimulates the growth of heterotrophic bacteria, resulting in deterministic and recurrent dynamics driven by their carbohydrate utilisation capacity [29, 58]. Phytoplankton production is also the primary source of other organic compounds, particularly DOS, such as DMSP, taurine and sulfoquinovose [59]. Although many bacteria harbour the capacity for inorganic sulfate assimilation, this process is energetically expensive. As such, using DOS compounds reduces energetic requirements and can additionally provide access to other nutrients, such as taurine that can act as a carbon, nitrogen and sulfur source. Additional genes, either directly or indirectly related to phytoplankton production and degradation, were also enriched, such as methylamine.

Functions enriched under high-ice cover

50% fewer genes were enriched under high ice, and they were mostly restricted to the recycling of bacterial cell wall carbohydrates, proteins, amino acids, aromatics and ketone compounds (Figure 7). Under high-ice cover, at the depth of sampling in this study (70 – 80 m), phytoplankton are less prominent [23], which would limit the availability of fresh labile organic matter and necessitate alternative growth strategies. For instance, the enrichment of a nitrate reductase gene indicates a specific potential to use inorganic substrates (Figure 7). Enrichment of GH families 109 and 18, related to peptidoglycan degradation, suggests recycling of bacterial cell wall components as carbon and energy sources. GH18 is also known to contain chitinases [60] and thus could indicate the degradation of chitin-rich materials such as carapaces and fecal pellets [61]. An increased reliance on bacterial-derived organic matter is further supported by the enrichment of D-amino acid degradation-related genes (Figure 7), as D-enantiomers of amino acids are largely derived from bacteria [62]. The enrichment in peptidases indicates that proteinaceous compounds play a more prominent role under high-ice, likely related to the production of related substrates by almost all organisms and hence wider availability.

Furthermore, we observed enrichment in genes for the degradation of aromatic and ketone compounds. Aromatic compounds in the Arctic Ocean typically originate from terrestrial organic matter that is sourced from rivers, constituting up to 33% of all Arctic Ocean DOM [63]. 12 – 41% of this terrestrial-derived DOM is exported to the North Atlantic via the EGC [63]. Therefore, the enrichment of genes for aromatic compound degradation indicates adaptations towards more diverse substrates under high-ice cover. These observations match reports for the *Chloroflexi* (SAR202) phylum inhabiting Arctic surface waters, which encode

more aromatic compound degradation genes compared to their deep-water counterparts [64]. Together, these features suggest the presence of a community that “recycles” available substrates and is not reliant on fresh labile-organic matter from phytoplankton.

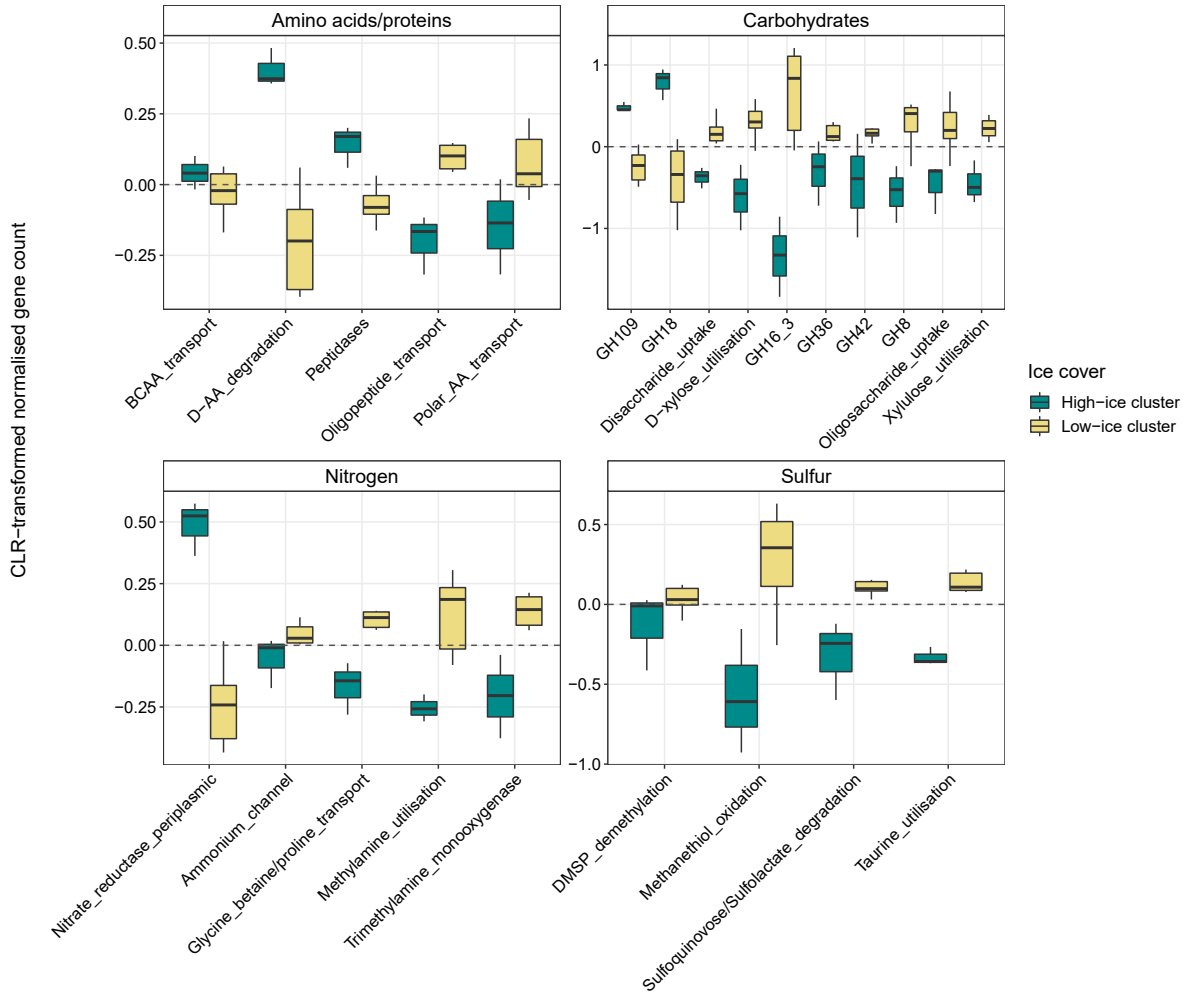


Figure 7. Selected genes involved in the uptake and degradation of organic and inorganic compounds enriched under high- and low-ice conditions. Enriched functional genes are displayed as the centered-log ratio transformed normalized gene counts. Where several genes of a single pathway were identified as enriched, they were grouped into one and the term 'utilization' used (e.g. "taurine_utilisation" indicates the uptake and degradation of taurine). When single genes were identified, the corresponding gene names are included.

Shifts in eukaryotic communities further support changes in ecosystem functioning

The observed ecosystem shifts between high and low-ice cover was reflected in the composition of eukaryotes at the same nine time points. High-ice cover conditions harbored increased proportions of *Dinophyceae*, *Syndiniales* and RAD-C radiolarians (Supplementary Figure S8), corresponding to prior reports of autumn-winter Arctic eukaryotic communities being dominated by heterotrophic-mixotrophic taxa [23, 65]. Although the proportions of diatoms (Bacillariophyta) were comparable at high- and low-ice, clear distinctions occurred at higher taxonomic resolution. High-ice conditions coincided with increased proportions of ice-

associated taxa such as *Bacillaria*, *Naviculales* and *Polarella*. In contrast, the open-water diatoms *Thalassiosira* and *Pseudo-nitzschia* as well as the prymnesiophyte *Phaeocystis* prevailed under low-ice cover, resembling temperate Atlantic phytoplankton communities (Supplementary Figure S8). Overall, the predominance of ice-associated algae and heterotrophic-mixotrophic eukaryotes under dense ice, compared to pelagic diatoms under low-ice cover, supports the distinction in ecosystem functioning relating to ice and daylight regimes.

CONCLUSION

Climate change is amplified in the Arctic Ocean, driving fundamental shifts in oceanographic and biological regimes. Here we show that variations in sea-ice extent and influx of Atlantic water masses considerably affect bacterial community dynamics and functionality. Densely ice-covered polar waters harbour a temporally stable, resident microbiome, which “recycles” available substrates and contains enriched signatures of autotrophic and inorganic substrate metabolism. In contrast, the ice margin with low-ice cover and more Atlantic water is dominated by seasonally fluctuating chemoheterotrophic populations, many of which are functionally linked to phytoplankton-derived organic matter. Both at population and community level, sea-ice cover had the strongest influence on bacterial functionality. Hence, we predict that the future Atlantification of the Arctic Ocean and continued reduction in sea-ice cover will shrink the ecological niches of signature Arctic populations, likely restricting them to the central basin and the core of the East Greenland current. These shifts represent a “Biological Atlantification” of the Arctic Ocean with implications for future ecosystem functioning and carbon cycling.

METHODS

Seawater collection and processing

Autonomous sample collection and subsequent processing of samples proceeded according to Wietz et al. [23]. Briefly, Remote Access Samplers (RAS; McLane) were deployed over four consecutive annual cycles between 2016-2020, with deployments and recovery occurring each summer (recover of 2019/2020 mooring occurred in 2021). From 2016 – 2018, RAS were deployed in the MIZ (78.83° N -2.79° E) and from 2018 – 2020 in the core EGC (79°N -5.4°E) at nominal depths of 80 and 70m, respectively. Sampling occurred at weekly to biweekly intervals (Supplementary Table S1). At each sampling event, ~1 L of seawater was autonomously pumped into sterile plastic bags and fixed with mercuric chloride (0.01% final concentration). After RAS recovery, water was filtered onto 0.22 µm Sterivex cartridges directly frozen at -20 °C until DNA extraction.

Amplicon sequencing and analysis

DNA was extracted using the DNeasy PowerWater kit (Qiagen, Germany), followed by amplification of 16S rRNA gene fragments using primers 515F–926R [66]. Sequencing was performed on a MiSeq platform (Illumina, CA, USA) using 2 x 300 bp paired-end libraries according to the "16S Metagenomic Sequencing Library Preparation protocol" (Illumina). Amplicons were subsequently processed into ASVs using DADA2. Analysis of ASV dynamics and subsequent generation of plots was performed in RStudio [67], using primarily, the vegan [68], limma [69], mixOmics [70], ggplot2 [71] and ComplexHeatmap [72] packages. Briefly, community composition was compared using Bray-Curtis dissimilarities and distance-based redundancy analysis with the functions *decostand* and *dbrda* in vegan and visualised using ggplot2. The influence of environmental variables on community dissimilarity was determined using the *ordiR2step* and *anova.cca* functions in vegan. ASV dynamics across the time-series and assignment into distribution groups, e.g. resident, was determined by extracting information from the relative abundance matrix produced by DADA2.

Co-occurrence networks were calculated for MIZ and core-EGC samples separately using the packages segmenTier [73] and igraph [74] in RStudio, and visualized in Cytoscape [75] with the Edge-weighted Spring-Embedded Layout. Nodes (ASVs) within the networks were only retained if the correlation coefficient was >0.7. Detailed pipeline is available in Supplementary Information S1.

PacBio metagenome sequencing

Nine samples from the 2016 – 2017 annual cycle at the MIZ were selected for metagenomic sequencing, using the same DNA as used for amplicon sequencing. Sequencing libraries were prepared following the protocol "Procedure & Checklist – Preparing HiFi SMRTbell® Libraries from Ultra-Low DNA Input" (PacBio, CA, USA) and subsequently inspected using a FEMTOpulse. Libraries were sequenced on 8M SMRT cells on a Sequel II platform for 30 h with sequencing chemistry 2.0 and binding kit 2.0. The sequencing was performed in conjunction with samples of another project, such that seven samples were multiplexed per SMRT cell. This resulted in, on average, 268 000 reads per metagenome, with an N50 of 6.8 kbp.

Taxonomic and functional annotation of HiFi reads

The 2.4 million generated HiFi reads were processed through a custom taxonomic classification and functional annotation pipeline. The classification pipeline followed similar steps to previously published tools, but with some modifications. A local database was constructed based on protein sequences from all species-representatives in the GTDB r202

database [76]. Prodigal v2.6.3 [77] was used to predict open reading frames (ORF) on HiFi reads, which were subsequently aligned to the GTDB-based database using Diamond blastp v2.0.14 [78] with the following parameters: --id 50 --query-cover 60 --top 5 --fast. After inspection of the hits, a second filtering step was performed: percentage identity of >65% and an e-value threshold of 1E-10. Using Taxonkit v0.10.1 [79], the last common ancestor (LCA) algorithm was performed, resulting in a single taxonomy for each ORF. A secondary LCA was subsequently performed for all ORFs from the same HiFi read, generating a single taxonomy for each read. Functional annotation of HiFi reads was performed using an extensive set of general and specialised databases. In brief, an initial gene annotation was performed using Prokka [80]. Then, a set of specialised databases were searched using blastp v2.11.0 [81] and HMMscan (HMMER v3.2.1) [82] for further gene annotations, including dbCAN v10 [83], CAZy (release 09242021) [84], SulfAtlas v1.3 [85], the Transporter Classification database [86], MEROPS [87], KEGG [88] and sets of Pfam HMM family profiles for SusD and TonB-dependent transporter genes. In order to compare functional gene composition across samples, gene counts were normalised by the average count of 16 universal, single-copy ribosomal proteins per sample [89] – providing 'per genome' counts.

Metagenome-assembled genome recovery

In order to maximise the recovery of MAGs, we clustered the metagenomes into two groups. The groups were determined based on the dissimilarity in ASV composition of the corresponding samples, and largely reflected samples of high- and low-ice cover. The samples were then individually assembled using metaFlye v2.8.3 (parameters: --meta --pacbio-hifi –keep-haplotypes --hifi-error 0.01). Contigs with a length of <10 kbp were removed and the remaining contigs were renamed to reflect the sample of origin. The resulting contigs from each group were concatenated into a single file. Coverage information, necessary for binning, was acquired through read recruitment of raw reads from all metagenomes to the contigs using Minimap2 v2.1 [90], using the 'map-hifi' preset. Contigs were binned using Vamb v3.0.2 [91] in multisplit mode using three different sets of parameters (set1: -l 32 -n 512 512, set2: -l 24 -n 384 384 and set3: -l 40 -n 768 768, as suggested by the authors). Completeness and contamination estimates of bins were determined using CheckM v1.1.3 [92] and those with >50% completeness and <10% contamination were manually refined using the Anvi'o interactive interface (Anvi'o v7) [93]. MAGs from both groups were combined and dereplicated at 99% average nucleotide identity using dRep v3.2.2 [94] (parameters: -comp 50 -con 5 -nc 0.50 -pa 0.85 -sa 0.98), resulting in 47 population-representative MAGs. A phylogenetic tree was reconstructed using the representative MAGs from this study and those recently published from the Fram Strait by Priest et al. [22] following a procedure outlined previously [89]. Briefly, 16 single-copy universal ribosomal proteins were identified in each MAG using

HMMsearch against the individual Pfam HMM family profiles and aligned using Muscle v3.8.15 [95]. The alignments were trimmed using TrimAl v1.4.1 [96], concatenated and provided as an input to FastTree v2.1.0 [97]. The tree was visualised and annotated in iToL [98].

Classification, abundance and distribution of MAGs

A dual taxonomic classification of MAGs was performed using single-copy marker and 16S rRNA genes. Firstly, MAGs were assigned a taxonomy using the GTDBtk tool v1.7.0 [99] with the GTDB r202 database. Secondly, extracted 16S rRNA gene sequences were imported into ARB [100], aligned with SINA [101] and phylogenetically placed into the SILVA SSU 138 Ref NR99 reference tree using ARB parsimony. Those containing a 16S rRNA gene were linked to ASV sequences through competitive read recruitment using BBMap of the BBtools program v35.14, with an identity threshold of 100%.

The distribution of MAGs across the nine metagenome samples generated for this study and an additional 42 Arctic metagenomes from the Tara Oceans collection (Project: PRJEB9740) was determined through read recruitment. Counts of competitively mapped reads were converted into the 80% truncated average sequencing depth, TAD80 [102]. Relative abundance was then determined as the quotient between the TAD80 and the average sequencing depth of 16 single copy ribosomal proteins. Ribosomal proteins were identified following the same procedure outlined above and their sequencing depth estimated using read recruitment with minimap2, for the metagenomes of this study, and BBMap, for Tara Oceans metagenomes.

Mooring and satellite data

To place bacterial community data into context, we incorporated a collection of in situ environmental parameters that are presented in Supplementary Table S1. Temperature, depth, salinity and oxygen concentration were measured using Seabird SBE37-ODO CTD sensors and chlorophyll *a* concentration was measured using a WET Labs ECO Triplet sensor attached to the RAS. Sensor measurements were averaged over 4 h around each sampling event. These parameters were subsequently used to determine the relative proportions of Atlantic Water (AW) and Polar Water (PW) as described previously by Wietz *et al.* [23]. Physical sensors were manufacturer-calibrated and processed in accordance with https://epic.awi.de/id/eprint/43137. Mooring-derived data are published under Pangaea accession 904565 [103] and 941159 [104]. Sea ice concentrations, derived from the AMSR-2 satellite, were downloaded from the https://seaice.uni-bremen.de/sea-ice-concentration-amsr-eamsr2, and averaged for the mooring regions using a 15 km radius.

DATA AND CODE AVAILABILITY

The 16S amplicon sequences are available at EBI-ENA under the project accessions PRJEB43890 (2016-17), PRJEB43889 (2017-18), PRJEB54562 (2018-19), PRJEB54586 (2019-20). Individual sample accessions are provided in Supplementary Table S5. The metagenomic sequence data and MAGs generated for this study are available at EBI-ENA under the project PRJEB52171. Supplementary table S6 contains the respective accession numbers for the individual metagenomic raw read datasets, assemblies and MAGs used in this study. Functional gene annotations for all signature populations are provided in Supplementary Files S1. Physicochemical parameters used in this study are available under the Pangaea accession 904565 [103] and 941159 [104].

The entire code for reproducing analyses and generating figures, along with necessary data files, are available under https://github.com/tpriest0/FRAM_EGC_2016_2020_data_analysis.

ACKNOWLEDGEMENTS

We thank Jana Bäger, Theresa Hargesheimer, Rafael Stiens and Lili Hufnagel for RAS operation; Daniel Scholz for RAS and sensor operations and programming; Normen Lochthofen, Janine Ludszuweit, Lennard Frommhold and Jonas Hagemann for mooring operation; Jakob Barz, Swantje Rogge and Anja Nicolaus for DNA extraction and library preparation, and Bruno Huettel and the technicians at the Max Planck Genome Centre in Cologne for metagenome sequencing. The captain, crew and scientists of RV Polarstern cruises PS99.2, PS107, PS114, PS121 and PS126 are gratefully acknowledged. Ship time was provided under grants AWI_PS99_00, AWI_PS107_05. We thank Oliver Ebenhöh and Eva-Maria Nöthig for helpful discussions. This project has received funding from the European Research Council (ERC) under the European Union's Seventh Framework Program (FP7/2007-2013) research project ABYSS (Grant Agreement no. 294757) to AB. Additional funding came from the Helmholtz Association, specifically for the FRAM infrastructure, and from the Max Planck Society.

AUTHOR CONTRIBUTION STATEMENT

TP performed ASV and metagenomics analysis. MW processed amplicon raw data into ASVs and coordinated the data analysis. TP and MW wrote the paper. WJvA contributed quality-controlled oceanographic data, and coordinated the mooring operations. EO and OP performed network analyses. STV provided quality-controlled chlorophyll sensor data. CB, KM and AB co-designed and coordinated the autonomous sampling and mooring strategy, and contributed to the interpretation of the results. BF and RA contributed to interpretation of results and development of the story. All authors contributed to the final manuscript.

REFERENCES

1. Kwok R. Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environ Res Lett*. 2018;13:105005.

2. AMAP. Snow, water, ice and permafrost in the Arctic (SWIPA). 2017. Oslo, Norway.

3. Notz D, Community S. Arctic sea ice in CMIP6. *Geophys Res Lett*. 2020;47:e2019GL086749.

4. Årthun M, Eldevik T, Smedsrud LH, Skagseth Ø, Ingvaldsen RB. Quantifying the influence of Atlantic heat on Barents sea ice variability and retreat. *J Clim*. 2012;25:4736–4743.

5. Polyakov IV, Pnyushkov AV, Alkire MB, Ashik IM, Baumann TM, Carmack EC, et al. Greater role for Atlantic inflows on sea-ice loss in the Eurasian Basin of the Arctic Ocean. *Science*. 2017;356:285–291.

6. Oziel L, Baudena A, Ardyna M, Massicotte P, Randelhoff A, Sallée J-B, et al. Faster Atlantic currents drive poleward expansion of temperate phytoplankton in the Arctic Ocean. *Nat Commun*. 2020;11:1–8.

7. Wassmann P. Arctic marine ecosystems in an era of rapid climate change. *Prog Oceanogr*. 2011;90:1–17.

8. Alonso-Sáez L, Sánchez O, Gasol JM, Balagué V, Pedrós-Alio C. Winter-to-summer changes in the composition and single-cell activity of near-surface Arctic prokaryotes. *Environ Microbiol*. 2008;10:2444–2454.

9. Underwood GJC, Michel C, Meisterhans G, Niemi A, Belzile C, Witt M, et al. Organic matter from Arctic sea-ice loss alters bacterial community structure and function. *Nat Clim Change*. 2019;9:170–176.

10. Arrigo KR, van Dijken GL. Continued increases in Arctic Ocean primary production. *Prog Oceanogr*. 2015;136:60–70.

11. Ardyna M, Arrigo KR. Phytoplankton dynamics in a changing Arctic Ocean. *Nat Clim Change*. 2020;10:892–903.

12. von Appen W-J, Waite AM, Bergmann M, Bienhold C, Boebel O, Bracher A, et al. Sea-ice derived meltwater stratification slows the biological carbon pump: results from continuous observations. *Nat Commun*. 2021;12:7309.

13. Rapp JZ, Fernández-Méndez M, Bienhold C, Boetius A. Effects of ice-algal aggregate export on the connectivity of bacterial communities in the Central Arctic Ocean. *Front Microbiol*. 2018;9:01035.

14. Fadeev E, Rogge A, Ramondenc S, Nöthig E-M, Wekerle C, Bienhold C, et al. Sea ice presence is linked to higher carbon export and vertical microbial connectivity in the Eurasian Arctic Ocean. *Commun Biol*. 2021;4:1–13.

15. Fadeev E, Wietz M, Appen W-J von, Iversen MH, Nöthig E-M, Engel A, et al. Submesoscale physicochemical dynamics directly shape bacterioplankton community structure in space and time. *Limnol Oceanogr*. 2021;66:2901–2913.

16. Nöthig E-M, Bracher A, Engel A, Metfies K, Niehoff B, Peeken I, et al. Summertime plankton ecology in Fram Strait—a compilation of long- and short-term observations. *Polar Res*. 2015;34:23349.

17. Soltwedel T, Bauerfeind E, Bergmann M, Bracher A, Budaeva N, Busch K, et al. Natural variability or anthropogenically-induced variation? Insights from 15 years of multidisciplinary observations at the arctic marine LTER site HAUSGARTEN. *Ecol Indic*. 2016;65:89–102.

18. Serreze MC, Barrett AP, Slater AG, Woodgate RA, Aagaard K, Lammers RB, et al. The large-scale freshwater cycle of the Arctic. *J Geophys Res Oceans*. 2006;111:C11010.

19. de Steur L, Hansen E, Mauritzen C, Beszczynska-Möller A, Fahrbach E. Impact of recirculation on the East Greenland Current in Fram Strait: results from moored current meter measurements between 1997 and 2009. *Deep Sea Res Part Oceanogr Res Pap*. 2014;92:26–40.

20. Hofmann Z, von Appen W-J, Wekerle C. Seasonal and mesoscale variability of the two Atlantic water recirculation pathways in Fram Strait. *J Geophys Res Oceans*. 2021;126:e2020JC017057.

21. Carter-Gates M, Balestreri C, Thorpe SE, Cottier F, Baylay A, Bibby TS, et al. Implications of increasing Atlantic influence for Arctic microbial community structure. *Sci Rep*. 2020;10:19262.

22. Priest T, Orellana LH, Huettel B, Fuchs BM, Amann R. Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms. *PeerJ*. 2021;9:e11721.

23. Wietz M, Bienhold C, Metfies K, Torres-Valdés S, von Appen W-J, Salter I, et al. The polar night shift: seasonal dynamics and drivers of Arctic Ocean microbiomes revealed by autonomous sampling. *ISME Commun*. 2021;1:1–12.

24. GEBCO Compilation Group. GEBCO 2020 Grid. 2020.

25. Galand PE, Potvin M, Casamayor EO, Lovejoy C. Hydrography shapes bacterial biogeography of the deep Arctic Ocean. *ISME J*. 2010;4:564–576.

26. Agogué H, Lamy D, Neal PR, Sogin ML, Herndl GJ. Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Mol Ecol*. 2011;20:258–274.

27. Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, et al. The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol*. 2009;11:3132–3139.

28. Auladell A, Barberán A, Logares R, Garcés E, Gasol JM, Ferrera I. Seasonal niche differentiation among closely related marine bacteria. *ISME J*. 2021;1–12.

29. Teeling H, Fuchs BM, Bennke CM, Krüger K, Chafee M, Kappelmann L, et al. Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife*. 2016;5:e11888.

30. Fadeev E, Salter I, Schourup-Kristensen V, Nöthig E-M, Metfies K, Engel A, et al. Microbial communities in the east and west Fram Strait during sea ice melting season. *Front Mar Sci*. 2018;5.

31. Eiler A, Hayakawa D, Rappé M. Non-random assembly of bacterioplankton communities in the subtropical North Pacific Ocean. *Front Microbiol*. 2011;2.

32. 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.

33. Kiene RP. Production of methanethiol from dimethylsulfoniopropionate in marine surface waters. *Mar Chem*. 1996;54:69–83.

34. Uhlig C, Damm E, Peeken I, Krumpen T, Rabe B, Korhonen M, et al. Sea ice and water mass influence dimethylsulfide concentrations in the Central Arctic Ocean. *Front Earth Sci*. 2019;7.

35. Park K-T, Lee K, Yoon Y-J, Lee H-W, Kim H-C, Lee B-Y, et al. Linking atmospheric dimethyl sulfide and the Arctic Ocean spring bloom. *Geophys Res Lett*. 2013;40:155–160.

36. Reisch C, Moran M, Whitman W. Bacterial catabolism of dimethylsulfoniopropionate (DMSP). *Front Microbiol*. 2011;2.

37. Vorholt JA, Chistoserdova L, Stolyar SM, Thauer RK, Lidstrom ME. Distribution of tetrahydromethanopterin-dependent enzymes in methylotrophic bacteria and phylogeny of methenyl tetrahydromethanopterin cyclohydrolases. *J Bacteriol*. 1999;181:5750–5757.

38. Gros V, Peeken I, Bluhm K, Zöllner E, Sarda-Esteve R, Bonsang B, et al. Carbon monoxide emissions by phytoplankton: evidence from laboratory experiments. *Environ Chem*. 2009;6:369–379.

39. Zuo Y, Jones RD. Formation of carbon monoxide by photolysis of dissolved marine organic material and its significance in the carbon cycling of the oceans. *Naturwissenschaften*. 1995;82:472–474.

40. Tolli JD, Sievert SM, Taylor CD. Unexpected diversity of bacteria capable of carbon monoxide oxidation in a coastal marine environment, and contribution of the Roseobacter-associated clade to total CO oxidation. *Appl Environ Microbiol*. 2006;72:1966–1973.

41. Hoarfrost A, Nayfach S, Ladau J, Yooseph S, Arnosti C, Dupont CL, et al. Global ecotypes in the ubiquitous marine clade SAR86. *ISME J*. 2020;14:178–188.

42. Francis B, Urich T, Mikolasch A, Teeling H, Amann R. North Sea spring bloom-associated Gammaproteobacteria fill diverse heterotrophic niches. *Environ Microbiome*. 2021;16:15.

43. Olson DK, Yoshizawa S, Boeuf D, Iwasaki W, DeLong EF. Proteorhodopsin variability and distribution in the North Pacific Subtropical Gyre. *ISME J*. 2018;12:1047–1060.

44. Yu Y, Yang J, Zheng L-Y, Sheng Q, Li C-Y, Wang M, et al. Diversity of D-amino acid utilizing Bacteria from Kongsfjorden, Arctic and the metabolic pathways for seven D-amino acids. *Front Microbiol*. 2020;10.

45. Pajares S. Unraveling the distribution patterns of bacterioplankton in a mesoscale cyclonic eddy confined to an oxygen-depleted basin. *Aquat Microb Ecol*. 2021;87:151–166.

46. Chun S-J, Cui Y, Baek SH, Ahn C-Y, Oh H-M. Seasonal succession of microbes in different size-fractions and their modular structures determined by both macro- and micro-environmental filtering in dynamic coastal waters. *Sci Total Environ*. 2021;784:147046.

47. Milici M, Vital M, Tomasch J, Badewien TH, Giebel H-A, Plumeier I, et al. Diversity and community composition of particle-associated and free-living bacteria in mesopelagic and bathypelagic Southern Ocean water masses: Evidence of dispersal limitation in the Bransfield Strait. *Limnol Oceanogr*. 2017;62:1080–1095.

48. Dalcin Martins P, de Jong A, Lenstra WK, van Helmond NAGM, Slomp CP, Jetten MSM, et al. Enrichment of novel Verrucomicrobia, Bacteroidetes, and Krumholzibacteria in an oxygen-limited methane- and iron-fed bioreactor inoculated with Bothnian Sea sediments. *MicrobiologyOpen*. 2021;10:e1175.

49. Lorenson TD, Greinert J, Coffin RB. Dissolved methane in the Beaufort Sea and the Arctic Ocean, 1992–2009; sources and atmospheric flux. *Limnol Oceanogr*. 2016;61:S300–S323.

50. Shakhova N, Semiletov I, Salyuk A, Yusupov V, Kosmach D, Gustafsson Ö. Extensive methane venting to the atmosphere from sediments of the East Siberian Arctic Shelf. *Science*. 2010;327:1246–1250.

51. Karl DM, Tilbrook BD. Production and transport of methane in oceanic particulate organic matter. *Nature*. 1994;368:732–734.

52. DeBruyn JM, Nixon LT, Fawaz MN, Johnson AM, Radosevich M. Global biogeography and quantitative seasonal dynamics of Gemmatimonadetes in soil. *Appl Environ Microbiol*. 2011;77:6295–6300.

53. Mujakić I, Piwosz K, Koblížek M. Phylum Gemmatimonadota and Its Role in the Environment. *Microorganisms*. 2022;10:151.

54. Benner R, Pakulski JD, Mccarthy M, Hedges JI, Hatcher PG. Bulk chemical characteristics of dissolved organic matter in the ocean. *Science*. 1992;255:1561–1564.

55. Lin P, Guo L. Spatial and vertical variability of dissolved carbohydrate species in the northern Gulf of Mexico following the Deepwater Horizon oil spill, 2010–2011. *Mar Chem*. 2015;174:13–25.

56. Panagiotopoulos C, Sempéré R. Analytical methods for the determination of sugars in marine samples: A historical perspective and future directions. *Limnol Oceanogr Methods*. 2005;3:419–454.

57. Kraan S. Algal polysaccharides, novel applications and outlook. Carbohydrates-comprehensive studies on glycobiology and glycotechnology. 2012. InTech, London.

58. 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*. 2012;336:608–611.

59. Ksionzek KB, Lechtenfeld OJ, McCallister SL, Schmitt-Kopplin P, Geuer JK, Geibert W, et al. Dissolved organic sulfur in the ocean: Biogeochemistry of a petagram inventory. *Science*. 2016;354:456–459.

60. Beier S, Bertilsson S. Bacterial chitin degradation—mechanisms and ecophysiological strategies. *Front Microbiol*. 2013;4.

61. Kirchner M. Microbial colonization of copepod body surfaces and chitin degradation in the sea. *Helgoländer Meeresunters*. 1995;49:201–212.

62. Radkov AD, Moe LA. Bacterial synthesis of d-amino acids. *Appl Microbiol Biotechnol*. 2014;98:5363–5374.

63. Opsahl S, Benner R, Amon RMW. Major flux of terrigenous dissolved organic matter through the Arctic Ocean. *Limnol Oceanogr*. 1999;44:2017–2023.

64. Colatriano D, Tran PQ, Guéguen C, Williams WJ, Lovejoy C, Walsh DA. Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. *Commun Biol*. 2018;1:90.

65. Freyria NJ, Joli N, Lovejoy C. A decadal perspective on north water microbial eukaryotes as Arctic Ocean sentinels. *Sci Rep*. 2021;11:8413.

66. Parada AE, Needham DM, Fuhrman JA. Every base matters: assessing small subunit rRNA primers for marine microbiomes with mock communities, time series and global field samples. *Environ Microbiol*. 2016;18:1403–1414.

67. RStudio Team. RStudio: Integrated development of R. 2015. RStudio Inc., Boston, MA.

68. Oksanen J, Blanchet FG, Friendly M, Kindt R, Legendre P, McGlinn D, et al. vegan community ecology package version 2.5-7 November 2020. 2020.

69. Ritchie ME, Phipson B, Wu D, Hu Y, Law CW, Shi W, et al. limma powers differential expression analyses for RNA-sequencing and microarray studies. *Nucleic Acids Res*. 2015;43:e47.

70. Le Cao K-A, Rohart F, Gonzalez I, Dejean S. mixOmics: Omics Data Integration Project. 2016.

71. Wickham H. ggplot2: Elegant Graphics for Data Analysis. 2016. Springer-Verlag New York.

72. Gu Z, Eils R, Schlesner M. Complex heatmaps reveal patterns and correlations in multidimensional genomic data. *J Bioinform*. 2016;32:2847–2849.

73. Machné R, Murray DB, Stadler PF. Similarity-based segmentation of multi-dimensional signals. *Sci Rep*. 2017;7:12355.

74. Csardi G, Nepusz T. The igraph software package for complex network research. *Interjournal Complex Syst*. 2006;1695.

75. Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, et al. Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Res*. 2003;13:2498–2504.

76. 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;50:785–794.

77. 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.

78. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12:59–60.

79. Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J Genet Genomics*. 2021;48:844–850.

80. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinforma Oxf Engl*. 2014;30:2068–2069.

81. Camacho C, Coulouris G, Avagyan V, Ma N, Papadopoulos J, Bealer K, et al. BLAST+: architecture and applications. *BMC Bioinformatics*. 2009;10:421.

82. Eddy SR. Accelerated Profile HMM Searches. *PLOS Comput Biol*. 2011;7:e1002195.

83. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2018;46:95–101.

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

85. Barbeyron T, Brillet-Guéguen L, Carré W, Carrière C, Caron C, Czjzek M, et al. Matching the diversity of sulfated biomolecules: creation of a classification database for sulfatases reflecting their substrate specificity. *PLOS ONE*. 2016;11:e0164846.

86. Saier MH Jr, Reddy VS, Moreno-Hagelsieb G, Hendargo KJ, Zhang Y, Iddamsetty V, et al. The Transporter Classification Database (TCDB): 2021 update. *Nucleic Acids Res*. 2021;49:D461–D467.

87. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. 2018;46:624–632.

88. Aramaki T, Blanc-Mathieu R, Endo H, Ohkubo K, Kanehisa M, Goto S, et al. KofamKOALA: KEGG Ortholog assignment based on profile HMM and adaptive score threshold. *J Bioinform*. 2020;36:2251–2252.

89. 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;1:1–6.

90. Li H. Minimap2: pairwise alignment for nucleotide sequences. *J Bioinform*. 2018;34:3094–3100.

91. Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol*. 2021;39:555–560.

92. 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.

93. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. 2015;3:e1319.

94. Olm MR, Brown CT, Brooks B, Banfield JF. dRep: a tool for fast and accurate genomic comparisons that enables improved genome recovery from metagenomes through de-replication. *ISME J*. 2017;11:2864–2868.

95. Edgar RC. MUSCLE: multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*. 2004;32:1792–1797.

96. Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *J Bioinform*. 2009;25:1972–1973.

97. Price MN, Dehal PS, Arkin AP. FastTree 2 – Approximately maximuml likelihood trees for large alignments. *PLOS ONE*. 2010;5:e9490.

98. Letunic I, Bork P. Interactive Tree Of Life (iTOL) v4: recent updates and new developments. *Nucleic Acids Res*. 2019;47:256–259.

99. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *J Bioinform*. 2020;36:1925–1927.

100. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: a software environment for sequence data. *Nucleic Acids Res*. 2004;32:1363–1371.

101. Pruesse E, Peplies J, Glöckner FO. SINA: accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinforma Oxf Engl*. 2012;28:1823–1829.

102. Orellana LH, Francis TB, Ferraro M, Hehemann J-H, Fuchs BM, Amann RI. Verrucomicrobiota are specialist consumers of sulfated methyl pentoses during diatom blooms. *ISME J*. 2021;1–12.

103. von Appen W-J. Physical oceanography and current meter data (including raw data) from FRAM moorings in the Fram Strait, 2016-2018. 2019. PANGAEA.

104. Hoppmann M, von Appen W-J, Monsees M, Lochthofen N, Bäger J, Behrendt A, et al. Raw physical oceanography, bio-optical and biogeochemical data from mooring HG-EGC-5 in the Fram Strait, July 2018 - August 2019. *PANGAEA*. 2022.

Supplementary Information S1

SUPPLEMENTARY RESULTS

Co-occurrence network analysis – methods

Co-occurrence networks were calculated for MIZ and core-EGC samples separately using the packages segmenTier[74] and igraph[75] in RStudio, and visualized in Cytoscape[76] with the Edge-weighted Spring-Embedded Layout. Oscillation signals were calculated for each ASV for each year based on discrete Fourier transformation of normalized abundances. Oscillation signals were then used to calculate the Pearson correlation between all ASVs, only retaining positive correlations. A network robustness analysis was performed to determine the minimal correlation value that represents a strong co-occurrence (0.7). Below this value, removal of a single node would cause network disruption. Full networks were then built including only above-threshold co-occurrences. Values of centrality and node betweenness were calculated using igraph.

Co-occurrence network analysis - results

Networks were computed for both moorings separately in order to assess patterns in microbial population dynamics under Arctic- and Atlantic-dominated conditions. In the core-EGC network, Res-ASVs exhibited twofold more significant co-occurrences than Int-ASVs, averaging 29 and 13 respectively (Figure 3b). In contrast, Int-ASVs were more connected in the MIZ network. This pattern further illustrates the stability of the resident microbiome under Arctic-dominated conditions. In both networks, a distinct cluster of co-occurring, summerly-peaking (June-August) ASVs were observed. These included ASVs reaching the highest relative abundances, such as asv24-*Luteolibacter* (Int-ASV), asv6-*Polaribacter* (Res-ASV) and asv7-*Aurantivirga* (Res-ASV). The MIZ network exhibited additional strong seasonal structuring. No further co-occurrence patterns in relation to environmental conditions were observed in the core-EGC network (Supplementary Figure S3).

Distribution of signature populations across the Arctic Ocean

In order to validate our observations on signature population dynamics with environmental conditions and their assignment as resident, intermittent and transient, we assessed their distribution across the Arctic Ocean using the Tara Oceans prokaryote size fraction dataset (Supplementary Figure S4). Signature population MAGs of the polar day clusters (C4 and C6) were, on average, present at the highest relative abundances in upper euphotic zone samples (5-100 m depth). In particular, asv6-*Polaribacter* in cluster C4, with an average relative abundance of 2.7%, and asv16-SAR92 Clade in cluster C6, with an average relative abundance of 2.1%. At lower depths (>100 m), polar day signatures decreased and Arctic and

polar night signature populations increased, i.e. as in C8 and C3, respectively. The most prominent were asv13-*Arenicellaceae* in cluster C3 and asv18-SAR86 Clade in cluster C8, with average relative abundances of 0.16 and 0.66% respectively. The Arctic Ocean sampling campaign of the Tara Oceans was conducted largely during summer months (May - October) and was restricted to locations above the continental shelf, which typically experience ice-free conditions or low-ice cover during that time period. As such, the higher prevalence of polar day signature populations (C4 and C6) in the upper water column is in agreement with their observed dynamics in the EGC. Water column stratification following sea-ice melt and polar day conditions likely restricts Arctic and polar night signature populations to deeper waters. These populations could be expected to increase in the upper water column in conjunction with deeper vertical mixing in the winter. Furthermore, it is likely that the "true" Arctic signature populations identified in our dataset are more prevalent in the Arctic Ocean basin, and are transported southward with water exiting the central Arctic through the EGC.

Ecological niches of signature populations

By combining temporal dynamics with functional gene predictions, we are able to make predictions on the ecological niches of signature populations within the context of the environmental conditions they represent. Signature populations of interest from clusters C1, C2 and C5 were characterized in detail in the main manuscript text, however additional signature populations from cluster C5 and those from C3, C4 and C6 were not described. Here, we provide ecological descriptions for the remaining signature populations.

Remaining Arctic signature populations (cluster C5)

The remaining Arctic signature populations were affiliated with *Nitrospina* (asv118) and the OM75 Clade (asv163), both of which exhibited similar dynamics and represented comparable proportions of the communities, reaching 1.7 – 1.8% relative abundance values. These populations harboured distinct functional features, relying on different substrates for growth but also performing key processes that contribute to the cycling of carbon, nitrogen and sulfur under high-ice coverage conditions.

The asv163 population was assigned to the OM75 Clade in the SILVA database and the GCA-2722775 (within the Nisaeaceae) in the GTDB database. Functional gene annotations indicated a non-motile lifestyle with a photoheterotrophic metabolism that included a green-light proteorhodopsin and the capacity to use a diverse repertoire of organic substrates for C, N and S acquisition and energy generation. Of particular note, was the extensive set of ABC transporters, 53, compared to other Arctic signature populations, which contained, on average, 17. Exogenous carbon substrates likely include osmolytes, C1 and aromatic compounds. The asv163 population encoded the C1 tetrahydrofolate oxidation pathway,

similar to some members of the SAR11 clade [1], which also produces energy through ATP. The degradative pathways for the osmolytes taurine, sarcosine and choline were encoded, whilst partial metabolic pathways for aromatic compound metabolism were also identified. Organic compounds also act as the main sulfur source for the asv163 population, with the degradation of sulfonates through 2-aminoacetaldehyde to sulfite. The sulfite could subsequently be converted to sulfate through the sulfite dehydrogenase gene (*soeABC*) and assimilated through 3'-phosphoadenylyl-sulfate. In addition to sulfonates, the asv163 population also harboured the capacity to utilise dimethylsulfide, dimethylsulfoxide and dimethylsulfone, which can act as a sulfur and carbon source, with the production of formaldehyde being channeled through the tetrahydrofolate oxidation pathway. Organic nitrogen sources for the asv163 population are predicted to include amino acids and urea. Several branched-chain and L-amino acid transporters were unique to this population as well as the complete urease enzyme complex. The asv163 population thus harbours an extensive capacity to uptake and degrade diverse organic substrates, which would be advantageous under high-ice conditions where the input of fresh organic matter is limited.

The asv118 population, assigned to the *Nitrospina* genus in SILVA and SCGCAAA288-L16 in GTDB, could be described as motile, chemolithotrophic organisms, in line with descriptions from previous studies [2,3]. *Nitrospina* belong to the *Nitrospinota* phylum that comprises the most abundant nitrite oxidising bacteria (NOB) in the oceans. Identified from surface to deep waters and from oxygenated to oxygen minimum zones, *Nitrospinota* are essential for the marine nitrogen cycle, with microbial nitrite oxidation reported to be the most significant biological pathway for nitrate production in the oceans[4]. However, the essential nitrite reductase, *nirK* gene, was not identified in the asv118 population, which may reflect the lower completeness of the MAG. The presence of a nitrite:ferredoxin reductase, *nirA*, indicates a capacity to convert nitrite to ammonia in an assimilation process and reflects previous obvious in other nitrite oxidizing bacteria of the *Nitrospira* genera[5]. Additional routes for nitrogen acquisition included ammonium uptake and a capacity to use urea through the *ureABC* gene. Additional observations in the asv118 population MAG included a green-light proteorhodopsin, suggesting supplemental energy generation through light - a feature not previously reported for *Nitrospina* members. Although the key nitrite oxidation machinery was lacking, we hypothesise that this process would still formulate a key part of energy generation in this organism, due to its highly conserved nature across NOB taxa. Recently, evidence for hydrogen oxidation in NOB taxa was reported, however no such genes were identified in the asv118 representative MAG.

Polar day signature populations (clusters C3 and C6)

Seven signature populations were identified for polar day conditions, six assigned to the Atlantic water-associated cluster C3 and one to the Arctic water-associated C6. Despite the affiliations to different sPLS clusters, C3 signature populations also reached comparable relative abundance values at the core-EGC under polar day conditions, suggesting that water mass is less influential. In contrast, the asv16 (SAR92 Clade) Arctic-water associated population did not reach comparably high relative abundances in MIZ samples, suggesting that this population is representative of polar day conditions only in Arctic water masses. Furthermore, the dynamics of the populations across the three polar day time periods in the core-EGC mooring appeared to be dependent on the magnitude of chlorophyll *a* concentrations measured; this pattern indicates an intrinsic link to phytoplankton dynamics and suggests that a certain threshold of phytoplankton abundance must be reached before such a response is observed in these populations. Each of the polar day signature populations identified here are affiliated with taxa that are well-known to be responders to phytoplankton blooms in marine environments [6,7]. At Helgoland Roads, a swift and recurrent succession of bacterial clades following phytoplankton blooms has been observed, with consecutive peaks of *Ulvibacter*, *Formosa, Reinekea, Polaribacter* and SAR92 from the *Bacteroidia* and *Gammaproteobacteria* class[8]. Here, the ephemeral peaks in relative abundance of some of these clades exhibited a more coordinated and less successional pattern, however this may reflect the lower resolution of sampling. In general, asv6 (*Polaribacter*), asv7 (*Aurantivirga*), asv55 (*Ulvibacter*), asv71/88 (*Nitrincolaceae*) and asv16 (SAR92) all exhibited rapid increases in relative abundance, peaking at the same time point followed by a decrease over a 4 week period. Although no chlorophyll *a* data is available from the MIZ mooring, the dynamics observed at core-EGC would indicate that these patterns are a response to phytoplankton blooms. As such, we hypothesise that these populations are involved in the degradation of phytoplankton-derived organic matter, but each occupying distinct substrate-based niches, as has been observed at Helgoland Roads.

Aurantivirga and *Polaribacter* have been shown to harbour broad substrate utilisation capacities but also occupy distinct polysaccharide-based niches[9,10]. In accordance with previous findings, both the asv6 (*Polaribacter*) and asv7 (*Aurantivirga*) representative MAGs harboured rich CAZyme gene repertoires and polysaccharide utilisation loci (PULs) for carbohydrate degradation. This consisted of 31 and 28 degradative CAZymes in asv6 and asv7, respectively, along with three distinct PULs in each. Two PULs identified in both of the populations are predicted to target the diatom storage polysaccharide laminarin (PUL 1 containing GH16_3, GH17 and GH149 and PUL 2 containing GH16_3 and GH30_1). The third PUL in the asv7 population is predicted to target α-glucans (GH13, GH13_31 and maltose transporter) whilst in asv6, sulphated xylose-containing polysaccharides are the predicted target for the third PUL (GH10, GH113, several sulfatases and D-xylose transporter). The

gene synteny and structures of the PULs are also in agreement with those previously described for *Polaribacter* and *Aurantivirga* representatives from Helgoland Roads[9]. Further comparisons revealed species-level differences in the dominant populations identified in our dataset and that of Helgoland Roads. For example, the asv6 Polaribacter population MAG shares 94.9% amino acid identity with 20120426_Bin_74_1 from Helgoland Roads (PRJEB28156), which was described as being present only in particular seasons and years but not related to the *Polaribacter* species that dominates during spring phytoplankton blooms. The above-described metabolic capacity, combined with ephemeral but pronounced peaks under polar day and alongside chlorophyll *a* peaks in the EGC, indicate the occupation of distinct substrate-based ecological niches for the asv6 and as7 populations. Furthermore, the comparable relative abundance values reached at both core-EGC and MIZ suggest that these populations are capable of proliferating under contrasting conditions, suggesting that substrate availability is the key factor defining their ecological niche.

In addition to the coordinated dynamics observed for members of the *Bacteroidia* and *Gammaproteobacteria*, three signature populations affiliated with the *Verrucomicrobiae* also exhibited pronounced increases in relative abundance. These populations were affiliated with the BACL24 (*Lentimonas*) and UBA1315 (*Luteolibacter*) genera. However, the dynamics of these populations differed. The asv94 population (*Lentimonas*) peaked with the *Bacteroidia* and *Gammaproteobacteria* representatives at core-EGC but showed a more delayed response at MIZ whilst the asv24 and asv115 (*Luteolibacter*) populations typically peaked later. These variations likely reflect the occupation of different ecological niches. Members of the *Verrucomicrobiae* are well evidenced to respond to phytoplankton blooms and are typically described as degraders of more complex polysaccharide structures, particularly those that are heavily sulfated[11,12]. In accordance with previous findings, the *Luteolibacter* representatives encoded a large number of degradative CAZymes, 40 in asv115 and 31 in asv24, and a high sulfatase to CAZyme ratio, 1:0.8 in asv115 and 1:0.7 in asv24. Further analysis on the encoded CAZyme genes revealed key distinctions between these two populations. Asv24 encoded genes assigned to five CAZyme gene families that are known to target alpha-glucans/amylose (GH13_38, GH13_4, GH13_8, GH57 and GH77) compared to only one gene in the asv115 (GH13_38). In addition, the asv115 population encoded two CAZymes that target rhamnogalacturonan (GH105 and GH106) which were absent from asv24. These metabolic distinctions may contribute to explaining the large difference in maximum relative abundances observed between these populations (6.1% in asv115 and 15.3% in asv24) and point towards substrate-based niche partitioning. Interestingly, the *Verrucomicrobiae* that are known to be the most prominent responders to spring phytoplankton blooms at Helgoland Roads are not from the *Luteolibacter* genera, but affiliated with different genera of the *Akkermansiaceae* family or with the *Lentimonas* genus of the *Puniceicoccaceae* family[12]. In

contrast, the *Lentimonas* population here (asv94) reached much lower relative abundances than the *Luteolibacter* populations. This further illustrates differences in the microbial populations that respond to phytoplankton blooms in different ecosystems.

Indifference to the above-described polar day-associated representatives, the asv71/88 population harboured distinct metabolic features, including a capacity for methylotrophy and a rich genetic repertoire for sulfur metabolism. Classified as *Nitrincolaceae* in the SILVA database and assigned to the ASP10-02a in GTDB, we describe the asv71/88 population as a motile chemoheterotroph. Methylotrophic metabolism was evidenced by genes involved in trimethylamine utilisation (*tmm, dmd-tmd, mgsABC* and *mbdAB*), with the produced formaldehyde likely being converting to CO2 through formate (*fdoG* and *fdwB*) and the ammonium being used as a nitrogen source. The asv71/88 population encoded a large number of genes involved in sulfur metabolism that included the ability to use organic sulfur compounds (methanesulfonate and sulfopyruvate) along with the complete thiosulfate oxidation machinery (*soxABCDXYZ*) and a sulfite dehydrogenase (*soeABC*). Methanesulfonate (MSA) is one of the main products of dimethylsulfide oxidation, and thus is likely present at higher concentrations during polar day conditions and phytoplankton blooms. The oxidation of MSA through a MSA monooxygenase, encoded in asv71/88 population, results in the production of formaldehyde and sulfite, which can be further oxidised through energy-generating reactions to CO2 and sulfate. Alongside the capacity to use organic nitrogen and sulfur compounds, the asv71/88 population also harboured 26 ABC transporters, including those for amino acid, monosaccharides, polyols and urea uptake, as well as the potential to degrade toluene. Furthermore, we identified the genes for a mannose-sensitive haemagglutinin-like pilus (*mshCDGIJKLOP*), which has been shown to promote attachment of bacterial cells to algae[13]. Therefore, the ability for motility and attachment combined with a diverse metabolic capacity of the asv71/88 population indicates a copiotrophic lifestyle that could involve a close relationship with phytoplankton cells, similar to that described for *Vibrio* and *Pseudoalteromonas* representatives. The encoded pathways for biotin, riboflavin, cobalamin and pantothenate synthesis, suggest that vitamins may be a valuable product provided to the phytoplankton by the asv71/88 population.

Polar night signature populations (clusters C4)

Signature populations of polar night conditions consisted of asv13, assigned to the *Arenicellaceae* family in SILVA and UBA11654 in GTDB, and asv8, assigned to the Arctic97B-4 in SILVA and UBA1096 in GTDB. The dynamics of these two populations were largely consistent, with highest relative abundance values of 4.6% for asv13 and 3.8% for asv8 reached under polar night conditions in MIZ samples. Insights into the metabolic capacity and potential ecological niches of these populations revealed some distinctions, however the

asv13 representative MAG was of lower completeness, 63%, which was reflected in the annotation of incomplete pathways that hindered the analysis.

Arenicellaceae is a family of Gammaproteobacteria that has previously been reported in deep-sea sediments[14], responding to phytoplankton blooms in coastal seawater[15] and has been proposed as an indicator of eutrophication[16]. The limited ecological information on members of this family indicates an organotrophic lifestyle. Due to the lower completeness of this MAG, we cannot provide clear predictions on the ecological niche but will briefly outline key metabolic features found. In general, the asv13 population could be characterized as non-motile and harbouring a capacity to use C1 compounds as substrates for growth along with indications of nitrate reduction (nitrate reductase, *narH*) and carbon fixation (incomplete rTCA cycle). C1 metabolism was indicated by the presence of a methanol dehydrogenase, formate dehydrogenase, methylenetetrahydrofolate dehydrogenase (*folD*) and the complete pathway for cofactor F420 biosynthesis. The presence of a nitrate reductase, *narH*, suggests a potential for nitrate reduction, however the other key subunits were missing. Two annotated ammonium channels also highlighted additional routes for nitrogen acquisition. We further focused on transport systems to reveal additional information on substrates used for growth, but those identified included typical transporters that are widespread in marine bacteria, such as vitamin B12, magnesium, sialic acid and general biopolymer transport (*exbBD*). Furthermore, we identified a complete riboflavin biosynthesis pathway. As a result, the ecological role of the asv13 population under polar night conditions is yet unknown and further analysis is required.

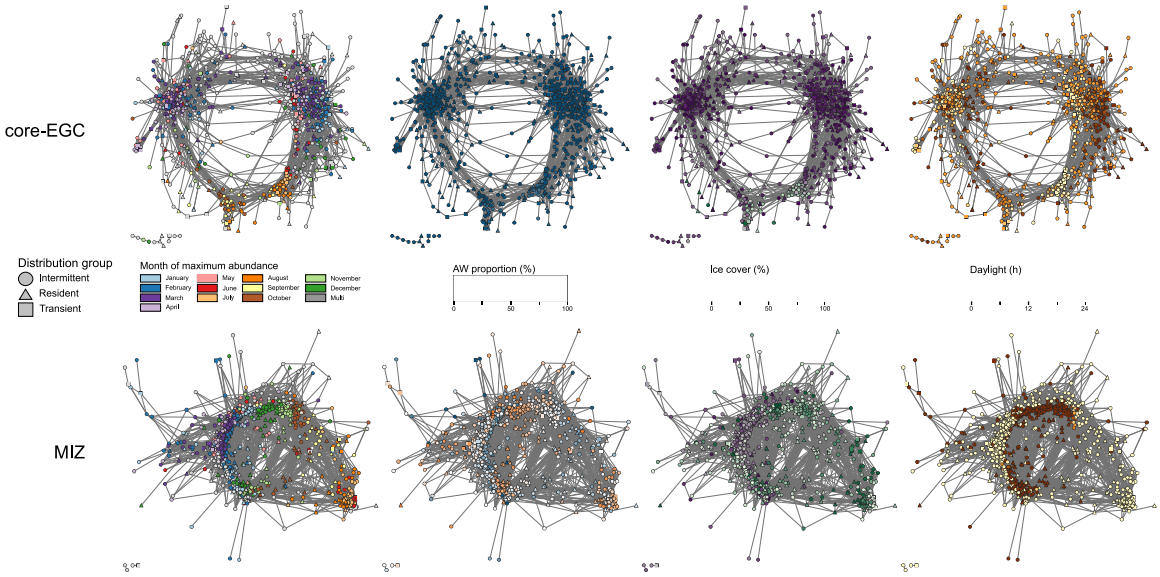
The asv8 population was assigned to the same taxonomic group in SILVA, Arctic97B-4, as the Arctic signature population (asv191), however the GTDB classification places these two populations into distinct families. This phylogenetic difference is also supported by distinctions in lifestyle and metabolic capacities. Indifference to the Arctic signature population, the asv8 population can be categorised as non-motile and harbouring an extensive genetic repertoire for organic compound metabolism, particularly carbohydrates. A total of 55 degradative CAZymes and 54 sulfatases were identified, highlighting a rich carbohydrate degradation capacity, which includes substrates such as pectin/rhamnogalacturonan (PL1 x 2, GH165 x 2, GH140 and CBM67 x 2), β-glucuronyl-containing polysaccharides (GH88, CE15 x 2) and sialic acids (GH33 x 3). Also encoded within the asv8 population was the capacity to use additional organic substrates as carbon and energy sources, which included glyoxylate and dicarboxylates, such as glycolate (*glcDEF*). For the acquisition of nitrogen and sulfur, inorganic sources are likely also used, with the complete sulfate assimilation pathway present and a nitrate reductase gene (*napA*) – although in the case of nitrate reduction not all necessary subunits and genes were present. Another key metabolic feature, which likely proves advantageous during winter conditions when fresh organic substrates are scarcely available, was the ability to synthesise glycogen (GBE1 and *glgACE*). Glycogen is a sugar

compound that plays important roles in energy and carbon storage in some bacteria and has been shown to increase bacterial durability under starvation conditions[17]. Although the internal hydrolysis of glycogen would not provide sufficient resources to explain observed dynamics during winter, it could certainly aid in population preservation under unfavourable conditions. With this, it could be hypothesized that the asv8 population is able to survive through the use of inorganic substrates and complex, recalcitrant carbohydrate substrates in winter conditions. The complex carbohydrates may be residual compounds left over from polar day conditions but may also be derived from under-ice algae and/or terrestrial-derived DOM.

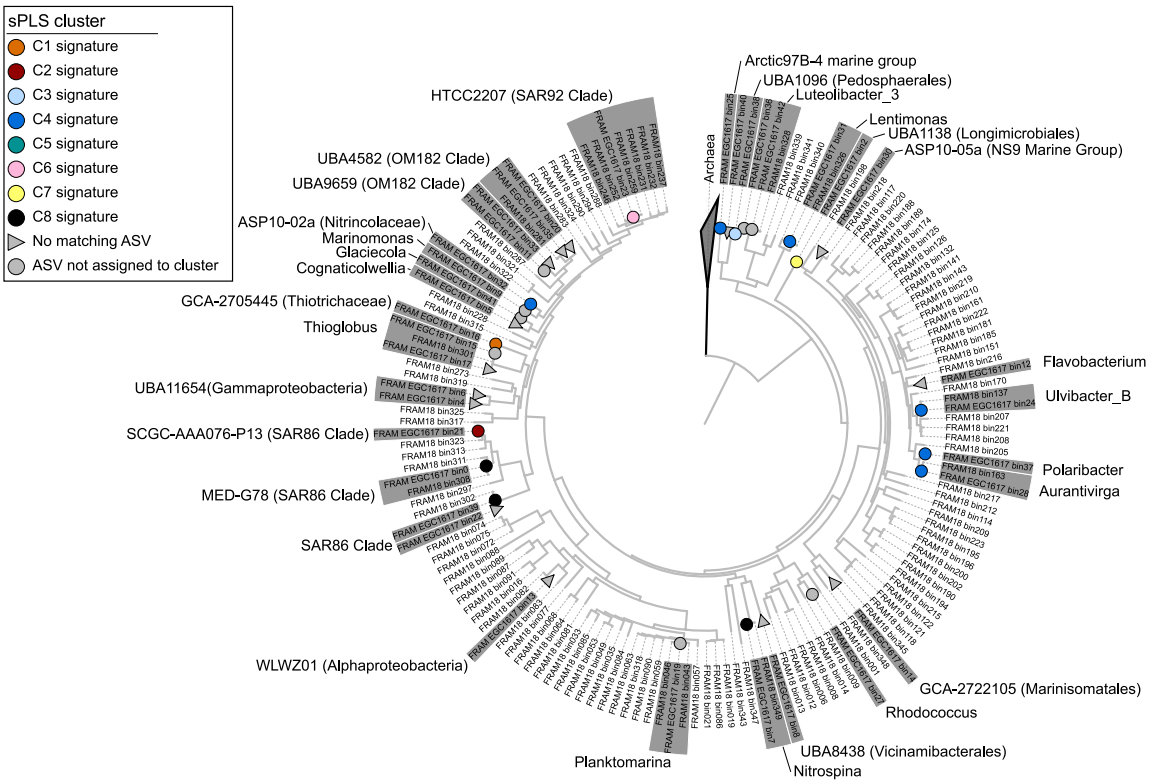
Supplementary tables and figures

All supplementary tables are available on the USB attached with this thesis and at the following link. Due to their large size, they were not included in the printed thesis. In addition, Supplementary Figure S1 and S2 are also not included in the printed thesis as they are animated GIFs.

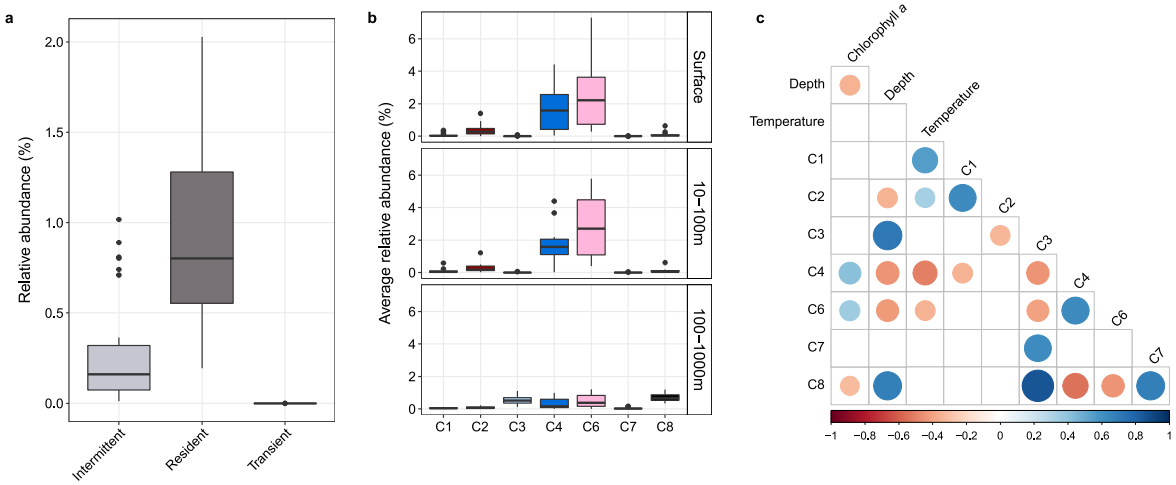
https://owncloud.mpi-bremen.de/index.php/s/O0U8Kj5fV6Bttmo



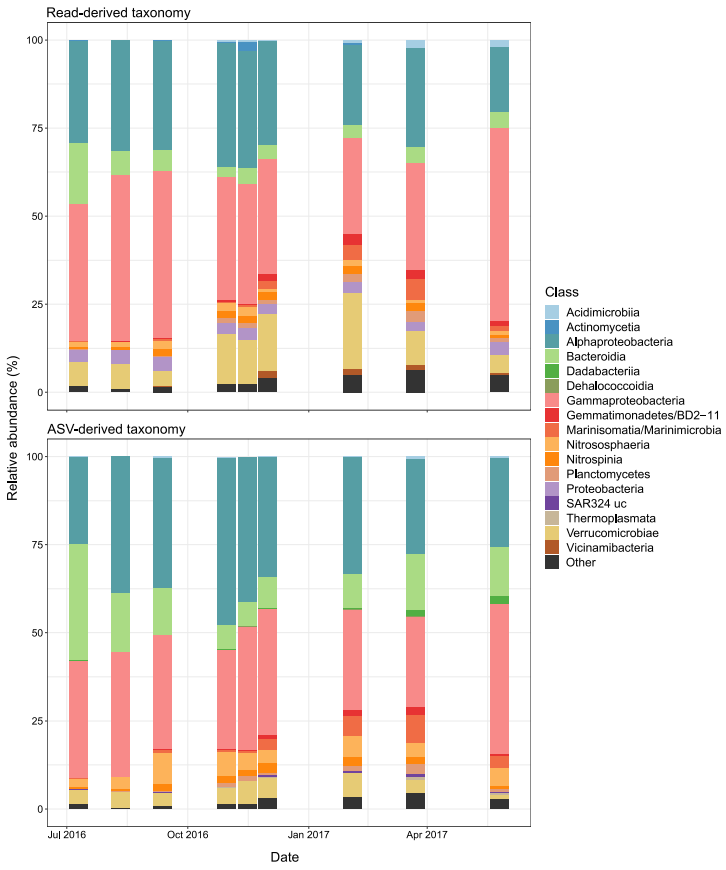
Supplementary Figure S3. Co-occurrence networks contextualised with metadata. Nodes in the network represent individual ASVs, with lines between nodes indicating co-occurence. The colouring of the nodes represents the time of year (a), AW proportion (b), Ice cover (c) and daylight (d) under which the ASV reached maximum relative abundance. Only co-occurences with a correlation value of >0.7 and a p-value of <0.05 were included in the network.



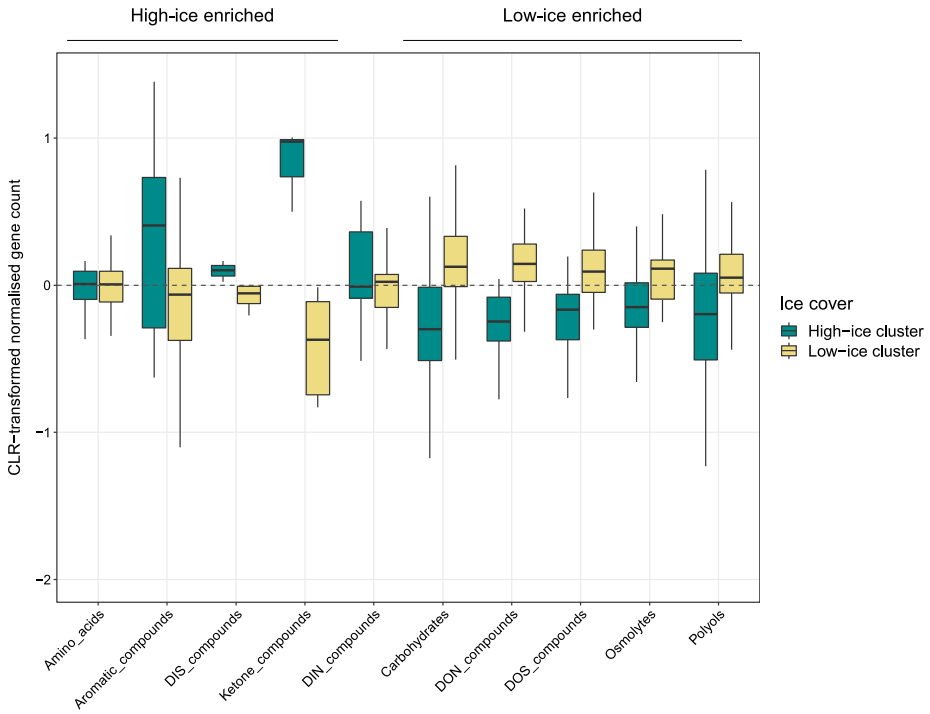
Supplementary Figure S4. Ribosomal protein-based phylogenetic tree of MAGs from this study and those previously recovered from the Fram Strait. Tree is based on concatenated alignment of 16 ribosomal proteins. MAGs from this study are indicated by coloured circles. Only MAGs that contained at least 8 ribosomal proteins were included in the tree. Grey boxes around lables indicate distinct genera that had MAG representatives recovered in this study.



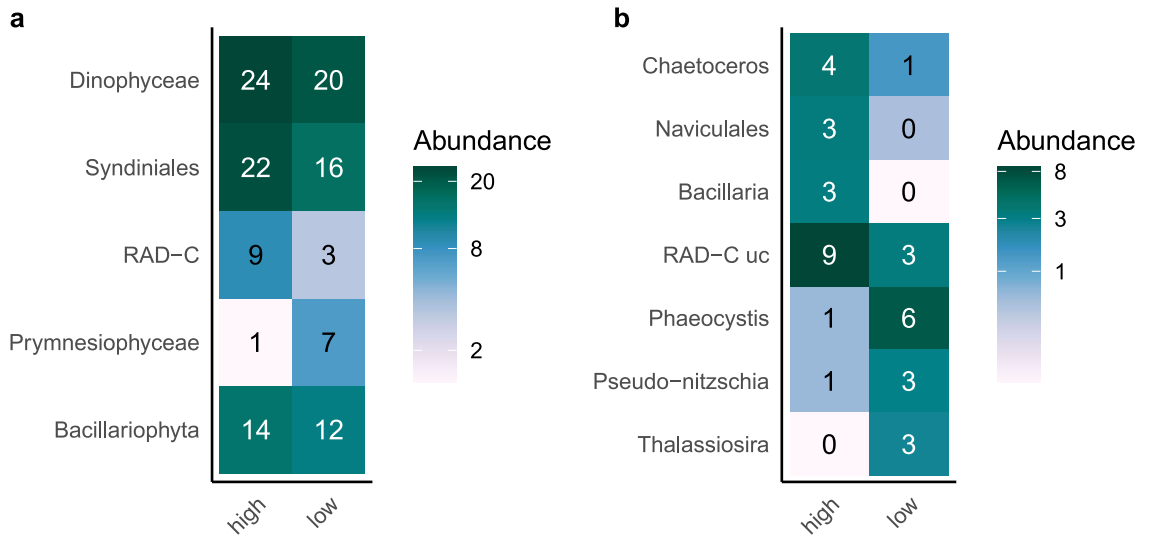
Supplementary Figure S4. Distribution dynamics of signature population MAGs across Tara Oceans Arctic metagenomes. Competitive read recruitment was performed using the prokaryote size fraction Tara Oceans Arctic metagenomes. Relative abundance was determined as the quotient of the truncated average depth of each MAG and the number of single copy ribosomal proteins in each metagenome. a) Average relative abundance of resident, intermittent and transient MAGs across all metagenomes. b) Average relative abundance of signature population MAGs in each sPLS cluster over the three distinct depth categories. c) Correlation of signature population MAGs in relation to each other and abiotic conditions of the metagenome samples. Only significant ($p < 0.05$) correlations were included in c).



Supplementary Figure S6. Comparison of PacBio HiFi read and ASV taxonomy at class-level. HiFi reads were taxonomically classified against a GTDB-based protein database. ASV taxonomy was assigned using the SILVA SSU 138 Ref NR99 database within the DADA2 pipeline.



Supplementary Figure S7. Summary of substrate uptake and degradation-related genes enriched under high and low ice cover. Enriched genes were individually searched in BioCyc and KEGG and subsequently grouped into broader categories based on the type of substrate metabolism they were involved in. Values shown are centered-log ratio transformed normalised gene counts, as used in the differential abundance analysis



Supplementary Figure S6. Heatmap illustrating the relative abundance of key eukaryotic taxa in high- and low-ice cover conditions. The high- and low-ice cover conditions correspond to the two distinct clusters identified when comparing whole microbial community functionality. The relative abundance of Eukaryotic taxa shown, are derived from the average relative abundance in the samples within each cluster.

Chapter V

Spatial heterogeneity in carbohydrates and their utilisation by microbial communities in the high North Atlantic

Taylor Priest, Silvia Vidal-Melgosa, Jan-Hendrik Hehemann, Rudolf Amann and Bernhard M. Fuchs

Manuscript in preparation

Contribution of the candidate in % of the total work

Experimental concept and design – 80%

Experimental work/acquisition of experimental data – 80%

Data analysis and interpretation – 80%

Preparation of figures and tables – 100%

Spatial heterogeneity in carbohydrates and their utilisation by microbial communities in the high North Atlantic

Taylor Priest[1]*, Silvia Vidal-Melgosa[1,2], Jan-Hendrik Hehemann[1,2], Rudolf Amann[1], Bernhard M. Fuchs[1]

[1] Max Planck Institute for Marine Microbiology, Bremen, Germany
[2] MARUM, University of Bremen, Bremen, Germany

***Corresponding author**
Taylor Priest
tpriest@mpi-bremen.de

ABSTRACT

Carbohydrates are structurally diverse, represent a substantial fraction of marine organic matter and are key substrates for marine heterotrophic microbes. Here, we combined analytical techniques with metagenomics and metatranscriptomics to characterise the carbohydrate fraction of particulate organic matter (POM) and the utilisation of carbohydrates by marine microbes in high latitude North Atlantic waters during late summer. The composition and abundance of monosaccharides and varied detection of specific polysaccharide epitopes revealed high spatial and vertical heterogeneity. Complex polysaccharides, known to accumulate in POM over time, were consistently detected, whilst more labile, simple structures, such as (1-3)-β-D-glucan, exhibited patchy distribution, indicating local variations in primary productivity. Metatranscriptomics revealed the presence of active and dynamic populations, with distinct assemblages dominating carbohydrate utilisation across samples. Variations in carbohydrate-active enzyme gene transcription revealed substrate-based niche partitioning among carbohydrate specialists, typically involving complex structures such as α-mannans and alginate, along with communal substrates, such as laminarin, targeted by multiple populations in each sample. The high spatial heterogeneity observed highlights how local biological and physical processes continue to shape the carbohydrate pool and the key microbial populations continue to be responsive during late summer in high latitude waters.

INTRODUCTION

Marine carbohydrates are structurally highly complex and represent a substantial fraction of characterised organic matter [1]. They are primarily produced by micro- and macroalgae to serve functions as cell wall components and as storage molecules for excess fixed carbon

and can constitute between 13 – 90% of algal biomass [2]. Carbohydrates are also prominent in phytoplankton exudates [3]. Phytoplankton-derived carbohydrates range from simple monosaccharides and low-molecular weight (LMW) oligosaccharides to complex high-molecular weight (HMW) polysaccharides, with the composition varying across taxa and with life stage and environmental conditions [4, 5]. Through exudation and upon cell death and lysis, a cocktail of LMW and HMW carbohydrates are released from phytoplankton and subsequently integrated into the dissolved and particulate organic matter pools. These compounds subsequently act as valuable substrates for heterotrophic microbes.

Exogenous carbohydrate utilisation is a widespread trait amongst bacterial and archaeal taxa however, the mechanisms employed and the types of carbohydrates that can be degraded vary [6–8]. Mono-, di- and trisaccharides can be readily taken up into the cell by many species through porins or dedicated transporters, whereas the uptake of oligosaccharides requires more specialised systems, such as TonB-dependent transporters (TBDTs). To make use of HMW polysaccharides, additional enzymes are needed. In particular, microbes must depolymerize the polysaccharide extracellularly using excreted or outer membrane-bound glycoside hydrolases (GHs) or polysaccharide lyases (PLs), often followed by uptake of the LMW products. These enzyme classes, together with carbohydrate-binding modules (CBMs) and carbohydrate esterases (CEs), are collectively referred to as carbohydrate-active enzymes (CAZymes). CAZymes are classified into families based on sequence similarity and common ancestry [9, 10], with many such families known to target specific glycosidic linkage types within polysaccharides. Some complex marine polysaccharides are decorated with other chemical groups, such as sulfates, which require additional genetic machinery for complete degradation, e.g. sulfatases. As such, comparing CAZyme gene profiles in conjunction with sulfatases and transporters can provide valuable insights into the carbohydrate utilisation potential of microbes [7, 11].

Carbohydrate utilisation by microbial populations exhibit spatial and temporal variations. The rate of hydrolysis and the substrate spectrum of extracellular CAZymes has been shown to reduce and narrow with depth [12] and with distance from the coast [13]. In addition, a latitudinal gradient has also been evidenced, with a broader spectrum of CAZyme activities measureable in temperate compared to high latitude waters [14]. Temporal shifts in CAZyme, sulfatase and transporter gene profiles are also evident following spring phytoplankton blooms [15]. These patterns are congruent with microbial population dynamics and shifts in microbial community composition. In particular, community-level patterns are shaped by the presence and composition of specialized carbohydrate degraders, such as members of the *Bacteroidetes. Bacteroidetes* typically harbour large CAZyme repertoires [7, 11] and exhibit distinct successional-like dynamics following spring phytoplankton blooms [16]. These dynamics have been described as evidence of substrate-based niche partitioning and

thus driven by the availability of substrate [15, 17]. Such detailed assessments on microbial carbohydrate utilisation have been largely focused on spring phytoplankton blooms in temperate ecosystems. Whether the same microbial clades occupy discrete substrate-based niches at later seasons and if these patterns are also evident in higher latitude waters is still yet to be determined.

In this study, we combine analytical techniques to characterise the carbohydrate fraction of particulate organic matter (POM) along with metagenomics and metatranscriptomics to assess microbial carbohydrate utilisation patterns in high latitudes of the North Atlantic during late summer. Through the application of an anion-exchange chromatography- and carbohydrate microarray-based approach, we were able to not only quantify particulate monosaccharides but also assess spatial and vertical distribution patterns of specific polysaccharide epitopes in the upper euphotic zone. Concurrently, using novel analysis pipelines on long PacBio HiFi sequences in concert with metatranscriptomics, we could characterise shifts in the potential and active carbohydrate utilisation patterns of entire microbial communities as well as specific populations.

RESULTS & DISCUSSION

Sampling of surface water (SW) and the bottom of the surface mixed layer (BML) was conducted at ten stations in the eastern Fram Strait and around Spitsbergen in late September 2020 (Figure 1). The stations could be further categorised as above-shelf, above-slope and open-ocean. This region is characterised by the West Spitsbergen Current (WSC) that transports Atlantic water northward, into the Arctic Ocean. The main branch of the WSC is topographically guided by the west Spitsbergen shelf break, which is typically characterised by a temperature-salinity front. On the west Spitsbergen shelf, Atlantic water (AW) converges and mixes with Arctic water and freshwater from land, resulting in intra-annual variability in hydrographic properties from a state of Arctic dominance in winter to Atlantic dominance in summer [18]. Previous studies have outlined typical thresholds that can be used to distinguish between the two main water masses in this region, which include >34.9 psu and >4.1 °C for AW [19]. The temperatures observed at SW and BML depths in our samples were indicative of AW, ranging from 4.1 – 7.7 °C, however, the salinity values of above-shelf (S10) and three above-slope stations (S1, S6 and S8) were below these thresholds (Figure 1 and Supplementary Table S1). These observations suggest an influence in the samples from these stations by polar and/or freshwater from Spitsbergen.

Monosaccharide and polysaccharide composition of samples

Quantifying the monosaccharide component of POM (>0.7 µm) in SW and BML samples revealed depth- and location-related patterns (Figure 2). Total monosaccharide

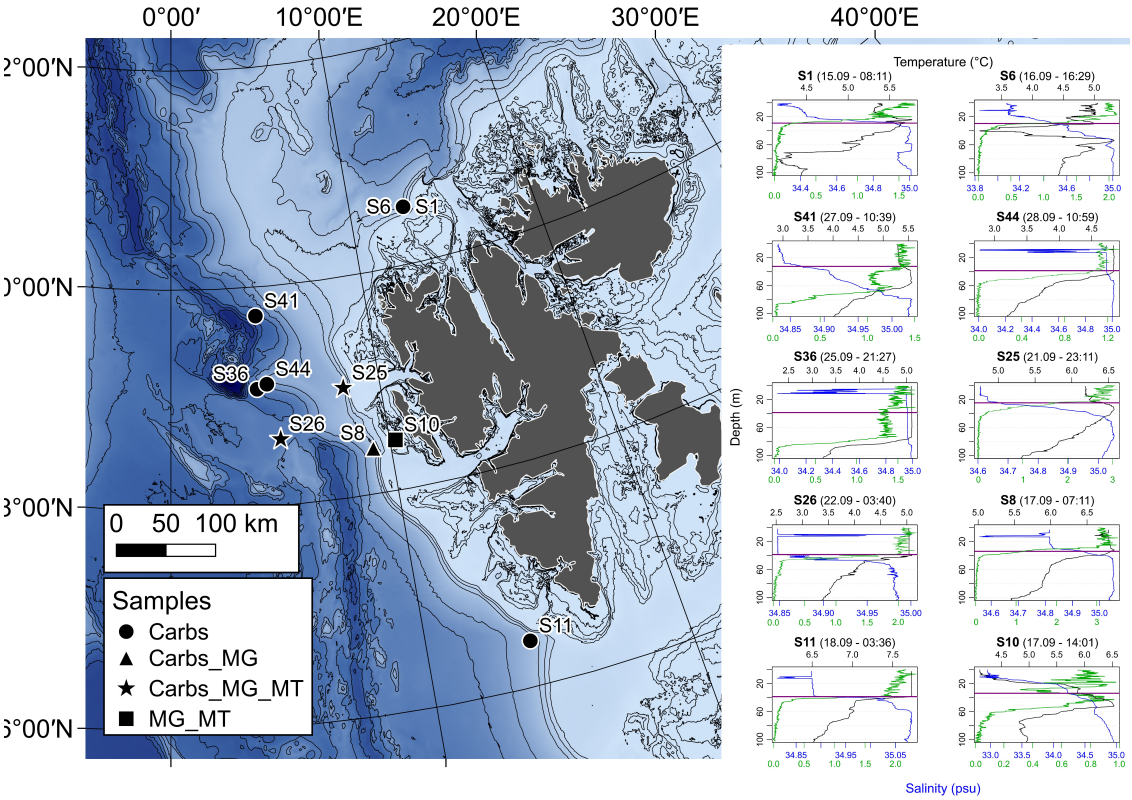


Figure 1. Bathymetric map with sampling locations, types of samples collected and vertical profiles from CTD casts. Map legend: Carbs = samples collected for carbohydrate analysis on particulate organic matter fraction, MG = sampled for microbial metagenomics, MT = sampled for microbial metatranscriptomics. Vertical profiles of temperature, salinity and fluorescence were derived from sensor measurements during CTD casts. The purple horizontal lines on each profile represent the bottom mixed layer (BML) sampling depth.

concentrations ranged from 0.5 – 5.2 µg l^{-1}, with higher concentrations typically observed in SW compared to BML samples, average of 2.9 µg l^{-1} and 1.7 µg l^{-1}, respectively (Supplementary Table S2). However, the magnitude of change between the two depths was station-dependent, from a negligible difference at station S8 to a threefold decrease at station S6 (Figure 2). These patterns are in agreement with previous findings that showed decreasing monosaccharide concentrations with depth, with the largest reduction occurring in the top 100 m of the water column [20]. In addition, stations categorised as above-slope typically contained higher total and individual monosaccharide concentrations than open-ocean stations (Figure 2 and Supplementary Figure S2). During early summer in this region, chlorophyll *a* and dissolved organic compounds have also been shown to reach higher concentrations in surface waters above the continental slope (~8 °E) [21]. These patterns likely reflect hydrographic processes, such as the frontal zone situated above the shelf break.

The most abundant monosaccharide detected in all samples was glucose, which represented a larger proportion of total monosaccharides in SW (47 – 51%) compared to BML (39 – 45%) samples and above-slope (44 – 51%) compared to open ocean (39 – 47%) samples (Figure 2). Previous studies have reported average glucose fractions in POM of 31 -

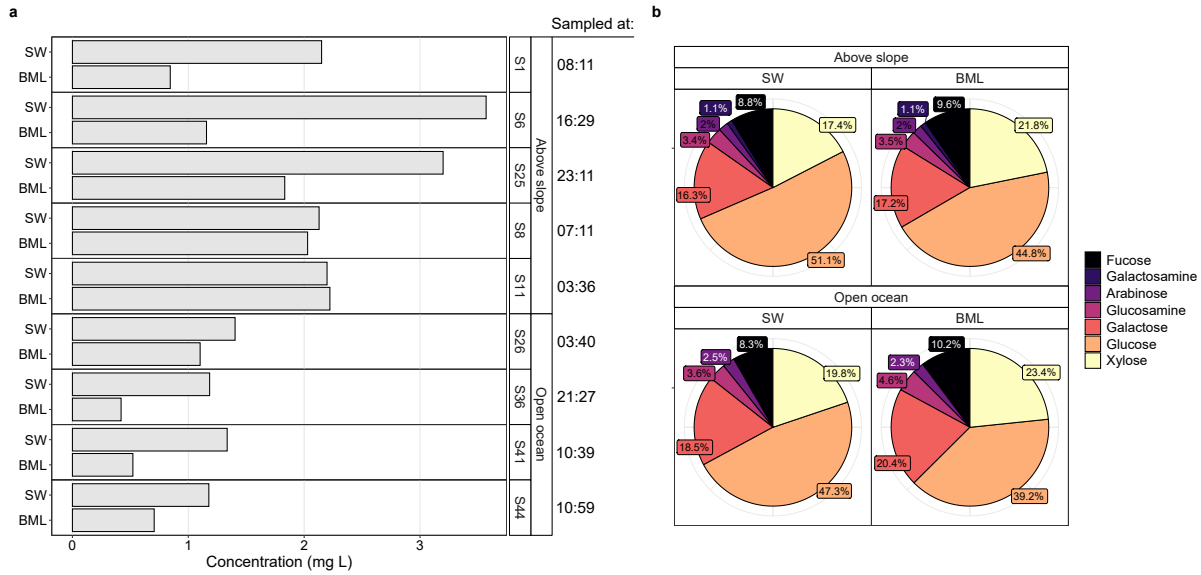


Figure 2. Total concentration and relative composition of monosaccharides in particulate organic matter fraction. a) Total monosaccharide concentrations at each station and depth. b) Relative composition of monosaccharides grouped by station location in relation ot continental slope. SW = surface water sample, BML = bottom of mixed layer.

55% in marine surface waters [22–24]. However, during phytoplankton blooms, the proportion of glucose has been shown to reach >70% [25]. Furthermore, in contrast to previous findings from the high North Pacific [22], we observed a decrease in glucose proportions with depth and an increase in all other monosaccharides. This pattern was most pronounced for xylose that, on average, increased by 5% in relative proportion from SW to BML depths (Figure 2). This indicates a preferential utilisation of glucose-based carbohydrates in the POM fraction in SW.

Using molecular probes combined with carbohydrate microarrays, we detected 17 distinct polysaccharide epitopes in POM (Supplementary Table S3). Variations in detection and abundance of epitopes revealed location and sample-specific patterns. The polysaccharide epitopes observed most frequently across all samples included arabinogalactan in 91% of samples, fucose-containing sulphated polysaccharides (FCSPs) in 86%, and 1-4-β-D-(galacto)(gluco)mannan in 77% (Figure 3). These widely detected epitopes represent complex polysaccharides that are known to constitute cell wall components and exudates of micro- and macroalgae [26]. In particular, FCSPs have been evidenced as the main carbohydrate excreted by diatoms in culture [27] and shown to accumulate in POM over a period of weeks following a spring phytoplankton bloom [28]. As such, the presence of these epitopes are likely signals of production that occurred earlier in the productive season. In contrast to this, (1-3)-β-D-glucan is a structurally simple polysaccharide (laminarin-like) that has previously been shown to exist only for days to a week in POM after a phytoplankton bloom [28]. (1-3)-β-D-glucan was not observed in all open-ocean samples and, on average, was detected at higher abundances in above-slopes (Supplementary Figure S2). In

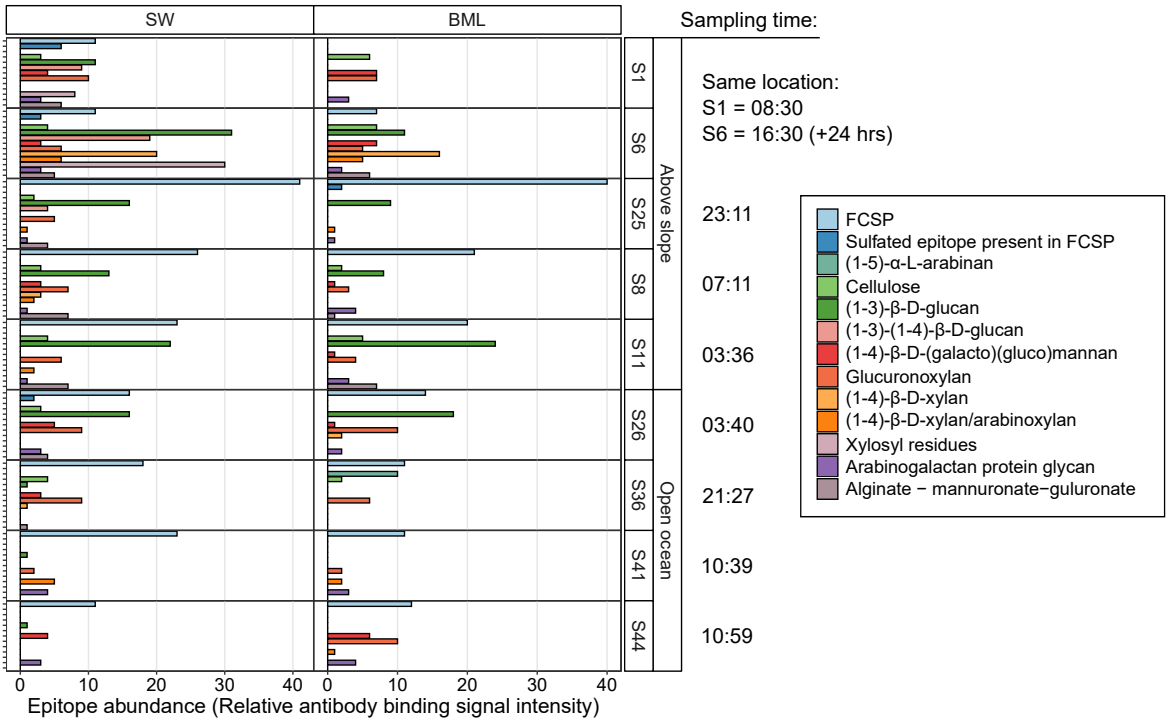


Figure 3. Composition of polysaccharide epitopes in the particulate organic matter fraction. Epitope abundance is a semi-quantitative value that represents the relative signal intensity of antibody binding in comparison to all other samples and standards. The epitope abundances incorporated here represent the summed values from the EDTA, MilliQ and NaOH extractions. SW = surface water sample, BML = bottom of surface mixed layer. FCSP = fucose-containing sulphated polysaccharide. Some epitopes that were only detected at low intensity in <3 samples were not included in this plot.

conjunction with the location-related differences in monosaccharide concentrations, open-ocean samples could be characterized as predominantly containing remnants of productivity from earlier seasonal stages whilst above-slope samples additionally feature more labile components, likely of recent production.

The collection of samples from the same location over a two day period (stations S1 and S6) and at additional depths (100 and 200 m) provided further insights into spatiotemporal changes (Supplementary Figure S3). S1 samples were retrieved in the morning whilst S6 samples were collected one day later in the evening. Temporal shifts in absolute monosaccharide concentrations were observed at all depths, with S6 samples containing 1.1 – 3.6 x higher concentrations than S1 samples along with the presence of arabinose that was not detected in S1 samples. Comparing relative monosaccharide compositions in SW samples also revealed several distinctions, including elevated proportions of glucose and galactosamine at S6 compared to S1, which was enriched in xylose, galactose and fucose. Clear temporal changes in polysaccharides were also observed, with higher signal intensity of (1-3)-β-D-glucan in S6_SW and the detection of (1-4)-β-D-xylan and (1-4)-β-D-arabinoxylan that were not detected in S1 samples. These observations, particularly the higher concentration of glucose and (1-3)-β-D-glucan in S6, indicate the addition of freshly produced polysaccharides between the sampling time points. Such temporal shifts could reflect diel

patterns in the particulate carbohydrate pool. Total particulate organic carbon typically shows strong diel periodicity, with accumulation during the day to a maximum concentration at dusk [29]. Within the particulate fraction, carbohydrates have also been shown to follow a similar pattern, with diurnal accumulation and nocturnal consumption [30].

Whole microbial community analysis

In order to assess microbial carbohydrate utilisation patterns, we generated eight PacBio HiFi read metagenomes from SW and BML depths from above-slope (S8 and S25) and open-ocean (S26) stations along with an above-shelf (S10) station, with four of the samples additionally used for metatranscriptomics. Details on generated metagenomes and metatranscriptomes are provided in Supplementary Table S4 and S5. Although monosaccharide and polysaccharide samples were not obtained at station S10, the physical characteristics of this station indicated an influence of Arctic and/or freshwater and thus we assumed it could have value in comparative analysis. An initial taxonomy-independent comparison was performed with previously published metagenomes from the Fram Strait and Arctic Ocean in order to place the sampled communities into context. Based on sequence composition dissimilarity, our metagenomes were most closely related to those previously generated from WSC, high North Atlantic and Barents Sea samples in June and July and most dissimilar to those sampled in the polar water mass of the East Greenland Current (Supplementary Figure S4). This indicates that the communities captured by the metagenomes are representative of summer communities in a North Atlantic water mass, also in agreement with the late summer composition of particulate carbohydrates.

As the focus of this study was on microbial carbohydrate utilisation, we initially classified reads at kingdom-level in order to remove eukaryotic sequences. Across the metagenomes, 22 – 49% of reads were identified as eukaryotic (Supplementary Figure 5a), which is surprising due to the size fractionated filtration used during sampling (see Methods). The analysis conducted hereon was performed only on those classified as prokaryotic in origin, however we extracted and analysed 18S rRNA genes to provide insights into the eukaryotic taxa present in each sample (Supplementary Figure S6). Using the average sequencing depth of single-copy ribosomal protein (SC-RBP) genes, we determined that the number of microbial genomes sequenced ranged from 761 in S8_SW to 1619 in S10_BML. Furthermore, using a subset of SC-RBPs with defined species-delineating average nucleotide identity (ANI) thresholds [31], we identified that the number of species per millions reads ranged from 374 in S8_SW to 554 in S10_BML (Supplementary Figure S5b). Combining these metrics resulted in 2 to 3.7 genomes per species per million reads being recovered, which suggests differences in species evenness across the samples.

Composition of metagenome and metatranscriptome microbial communities

The microbial diversity captured by the metagenomes consisted of 1644 species that formulated unique community structures in each sample, evident at the family and genus level. Phylogenetic composition of metagenome and metatranscriptome communities was compared using the large subunit ribosomal protein L6 (RBP L6), which has been described as highly recoverable and with the most accurate species-delineating ANI threshold of SC-RBPs [31]. Taxonomic classification was performed against GTDB however, if corresponding SILVA taxonomies are known then both will be stated upon first mention, followed by only the SILVA taxonomy – both classification schemes will be used in all figures. At the family-level, communities were dominated by *Pelagibacteraceae* (12 – 15%), *Rhodobacteraceae* (4 – 12%), D2472 (SAR86; 6 – 11%) and the *Flavobacteriaceae* (7 – 12%) except for in S25_BML, which was dominated by *Alteromonadaceae* (25%) (Supplementary Figure S7). The prominent genera identified across samples resemble those that are known to be dominant in summer microbial communities in the WSC [32, 33]. However, in difference to our observations from September 2020, a previous report on microbial communities in this region from late September 2018, indicated autumn-like compositions, which suggests high inter-annual variability [34].

Thirteen genera were identified as reaching >2.5% relative abundance and together represented, on average, 42% of the microbial communities. These included *Pseudoalteromonas,* MGIIa-L1, SAR86, ASP10-02a (*Nitrincolaceae*) and HTCC2207 (SAR92). The high proportions of MGIIa-L1 has not been previously described in the WSC, which may reflect differences in primers used and biases in 16S rRNA gene-based analyses. Applying the same >2.5% threshold to the relative proportion of transcription in the four metatranscriptome samples, resulted in 18 genera being identified that accounted for, on average, 69% of genome transcription in the microbial communities (Figure 4). Among those contributing the most to transcription was *Pseudoalteromonas, Vibrio, Pelagibacter* and *Candidatus* Marisimplicoccus [35] (NS5). Discrepancies between the largest contributing genera to metagenomic and metatranscriptome-derived communities were apparent, highlighting that higher relative abundance does not necessarily correspond to a higher proportion of transcription and vice versa. Two examples of this include the MGIIa-L1 that constituted 16% relative abundance in S10_SW but only 9% relative proportion of transcription and MAG-121220-bin8 (NS4 Marine Group) that comprised 2% relative abundance in S25_SW but 5% relative proportion of transcription. Such discrepancies are not uncommon in marine microbial communities [36] and can be influenced by lifestyle, fitness and differences in metabolism and substrate utilisation.

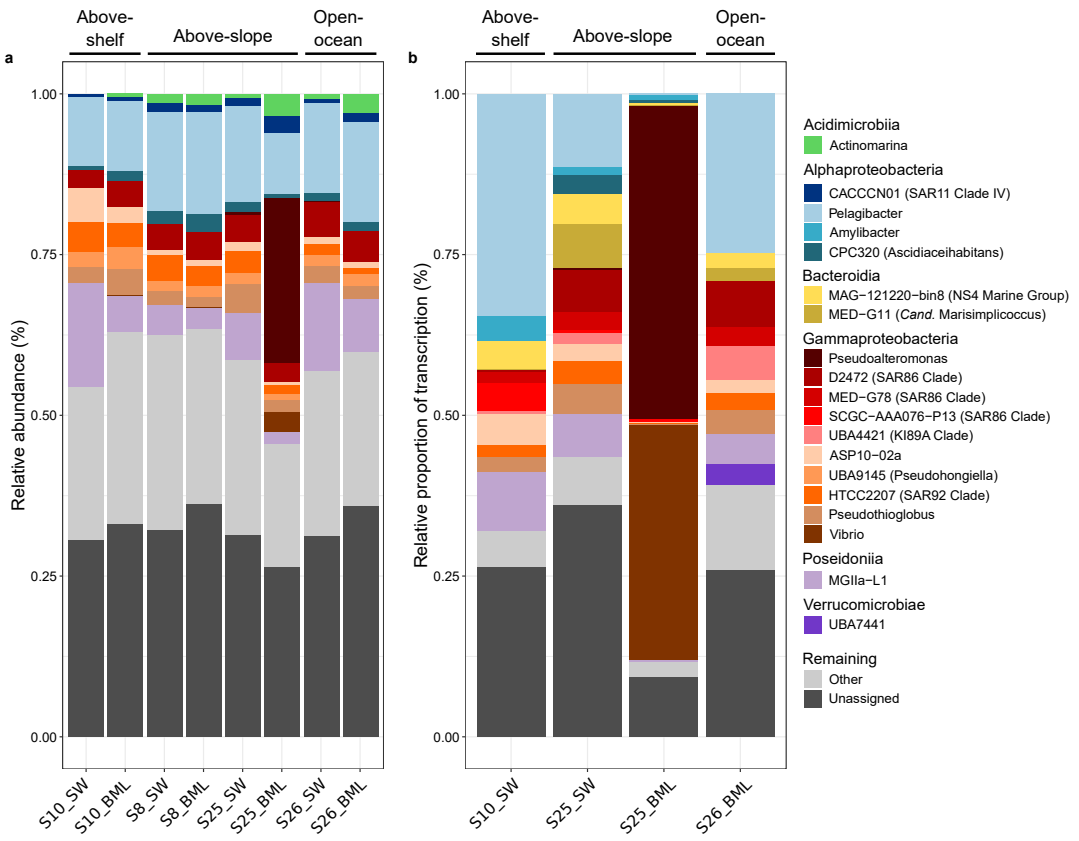


Figure 4. Genus level composition and transcription of sampled microbial communities based on ribosomal protein L6 gene. Relative abundance for each taxa is the quotient between the ribosomal protein L6 gene count and total sample ribosomal protein L6 gene count. Relative proportion of transcription is determined using the same approach as for relative abundance except TPM values were used. Taxonomy of ribosomal protein L6 gene sequences was derived from taxonomic classification of the original HiFi read against a GTDB database.

Although unique community structures were evident in all samples, sample S25_BML was the most distinct, with a community dominated by *Pseudoalteromonas* in the metagenome, and *Pseudoalteromonas* and *Vibrio* in the metatranscriptome (Figure 4). These genera have previously been observed under nutrient rich conditions and are known to be associated with eukaryotic hosts and phytoplankton blooms [37, 38]. Considering the high proportion of eukaryotic reads captured in the metagenomes, the observed pattern in S25_BML may represent signals of processes occurring in the larger size fraction. In support of this, the eukaryotic community of this sample (Supplementary Figure S6) contained a large proportion of 18S rRNA genes affiliated with copepods, whose microbiomes are typically dominated by *Pseudoalteromonas* and *Vibrio* [39, 40].

Whole community carbohydrate utilisation patterns

Carbohydrate utilisation patterns of the microbial communities in the metagenomes and metatranscriptomes were assessed through degradative CAZymes. In order to compare CAZyme abundances across samples, gene counts were normalized by the number of

genomes captured in each sample and will be referred to as counts per microbial genome (CPMG). CAZymes represented a low proportion of total recovered genes, 0.3% on average, which corresponded to 11 CPMG, on average (Supplementary Table S6). Comparing CAZyme gene compositions revealed no consistent patterns in relation to depth or location, e.g. above-slope vs open-ocean. However, similar compositions were observed within stations for S8 and S26, whilst larger variabilities were observed in S10 and S25 samples, which reflected differences in whole community functionality (Supplementary Figure S8).

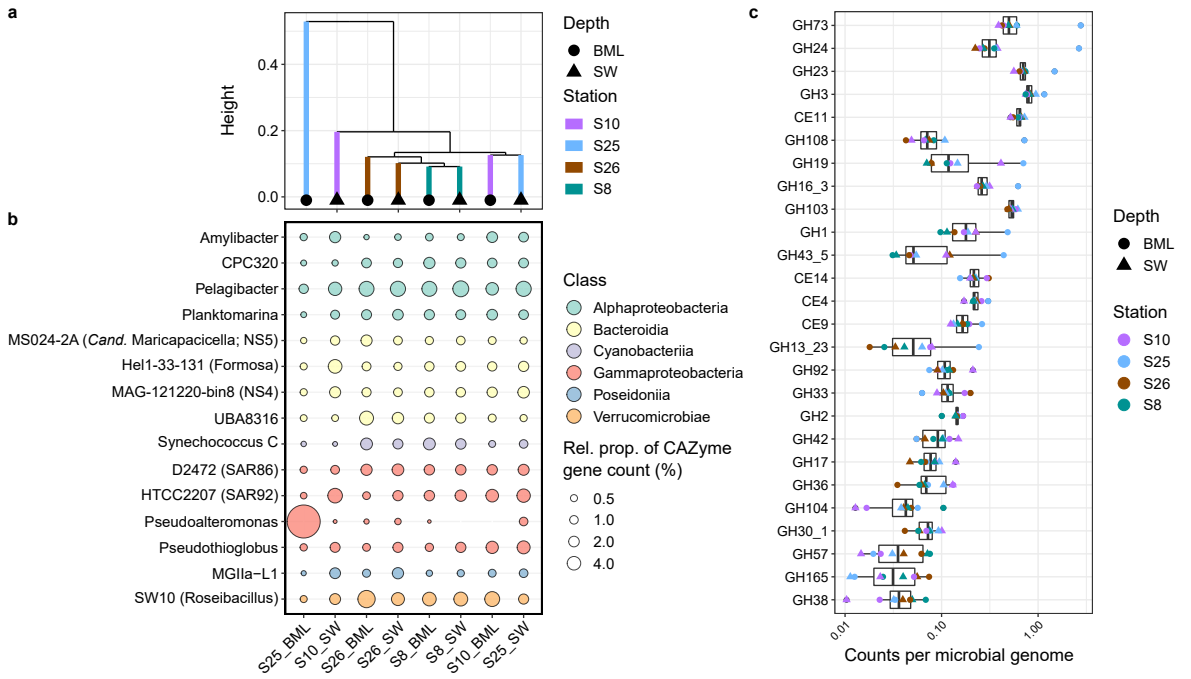


Figure 5. Composition, variation and taxonomic affiliation of degradative CAZymes across metagenome samples. a) Comparison of whole community carbohydrate-active enzyme (CAZyme) gene composition based on Bray-Curtis dissimilarities. b) Phylogenetic genera contributing the largest proportions to whole community CAZyme gene counts in each sample. c) CAZyme gene (sub-)families that exhibited the largest range in normalised gene count at the community level, across the samples.

The CAZyme gene families with the highest CPMG values were typically those related to peptidoglycan synthesis and degradation, including GH23, GH73, GH103 and GH3, which is in agreement with previous observations and reflects the core machinery required for cell membrane construction and maintenance [16]. We further identified 26 CAZyme gene families that exhibited a change in CPMG of >0.05 across samples (Figure 5c). Amongst these, unique sample-specific signatures were observed, including higher CPMG values for genes involved in peptidoglycan recycling in S25_BML (GH73, GH23 and GH108), degradation of β-1,3-glucan (GH17) and α-mannan (GH92) in S10_SW and β-galactose-containing polysaccharides (GH165) in S26_BML. Concurrently, changes in the taxonomic composition of CAZyme genes were also observed across samples, with a higher proportion being affiliated with *Amylibacter*, Hel1-33-131 (*Formosa*) and SAR92 Clade in S10_SW and UBA8316 (*Bacteroidia*) and SW10 (*Roseibacillus*) in S26_BML (Figure 5b).

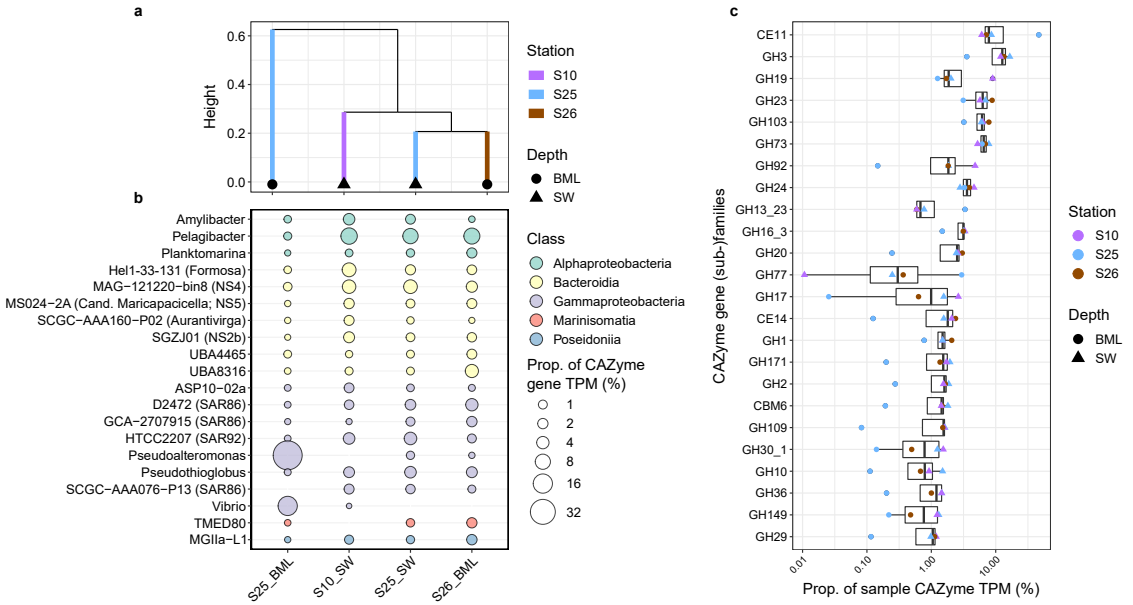


Figure 6. Composition, variation and taxonomic affilication of transcribed degradative CAZymes. a) Comparison of whole community transcribed carbohydrate-active enzyme (CAZyme) gene composition based on Bray-Curtis dissimilarities. b) Phylogenetic genera contributing the largest proportions to whole community CAZyme transcription in each sample. c) CAZyme gene (sub-)families that exhibited the largest range in TPM values at the community level, across the samples.

Sample-specific patterns in CAZyme gene profiles and the main contributing taxa were more pronounced in the metatranscriptomes (Figure 6). The proportion of total gene transcription affiliated with CAZymes differed considerably across the samples, from 1.5% in S25_BML to 3% in S10_SW. Further differences were observed in the number of transcribed CAZyme genes, with a much higher number in S10_SW, 3159 genes, compared to S25_BML, 494 genes (Supplementary Table S7). Comparing the CPMG and proportion of transcription of each CAZyme gene family revealed, in general, strong positive linear trends (Supplementary Figure S9) – higher CPMG resulted in higher proportion of transcription. However, several evident deviations from this trend were observed. Notably, CE11 in S25_BML (47%) and in S10_SW, GH19 (9%), GH92 (5%) and GH3 (12%). These families also constituted the sample-specific signatures for these samples, with higher proportion of transcription compared to other samples. Additional signatures were revealed by comparing the proportion of transcription of CAZyme families across samples (Figure 6c). These included two families in S25_BML that are known to target α-glucans, GH13_23 (2.6%) and GH77 (1.5%), a β-xylosidase (GH10) in S25_SW and higher proportions of several β-1,3-glucan-targeting families in both SW samples (GH30_1, GH17 and GH149). In some cases, these unique signatures could be attributed to specific taxonomic groups, such as the CE11 in S25_BML that was almost entirely associated with *Pseudoalteromonas*. Although CE11 is amongst the degradative CAZymes, its so far known function is related to the synthesis of lipid A, a key component of lipopolysaccharides.

The composition of taxa contributing the most to CAZyme transcription was unique in each sample, but was dominated by genera within *Bacteroidia* and *Gammaproteobacteria* (Figure 6b). The dominant *Bacteroidia* genera identified here are well known specialist degraders of carbohydrates in marine environments and, together with *Gammaproteobacteria* genera, constitute some of the main responders to phytoplankton blooms in coastal temperate ecosystems [16, 41]. More specifically, several of these genera are amongst those that exhibit successional-like dynamics and have been described as occupying discrete substrate niches [7, 16]. Here, we show that the same key genera are also dominating carbohydrate degradation in later seasonal stages. Furthermore, in addition to the previously described temporal dynamics, the above-described patterns highlight spatial heterogeneity in microbial community carbohydrate utilisation, which is largely driven by the dynamics of these specialist populations.

To provide higher resolution insights into spatial distribution patterns of carbohydrate utilisation and assess if niche partitioning is evident for the key carbohydrate-degrading populations, we investigated patterns at the population-level using metagenome-assembled genomes.

Population-level carbohydrate utilisation patterns

We recovered and analysed 83 population-representative metagenome-assembled genomes (MAGs), delineated at a 99% ANI threshold. Detailed information on the MAGs is provided in Supplementary Information S1 and Supplementary Table S8. The MAGs captured between 48 – 88% of the microbial metagenomic reads (Supplementary Table S9) and 11 – 37% of the microbial metatranscriptomic reads (Supplementary Table S10), and each exhibited variable abundances across the different metagenomes and metatranscriptomes (Figure 7). One MAG recovered, S25_BML_bin_129, shares 97.1% ANI to the cultured representative genome of *Pseudoalteromonas primoryensis,* and constituted 22.8% relative abundance of sample S25_BML. This indicates that the patterns observed at the community-level are driven by a single species. Similar to the patterns described in the community-level analysis, varying degrees of coupling between relative abundance and proportion of transcription of MAGs were observed and largely reflected taxonomic differences. For example, MGIIa-L1-affiliated MAGs exhibited high relative abundance but a low proportion of transcription whilst *Bacteroidia*-affiliated MAGs showed the opposite trend. The MGIIa-L1 MAGs were consistently among the most abundant in all metagenome samples. A recent study describing the ecology of MGII representatives in the German Bight, identified MGIIa-L1 as being indicative of post phytoplankton bloom situations and distinct from the dominant populations during winter that were affiliated with the MGIIb [42]. This observation thus further supports our sampling of a late summer community.

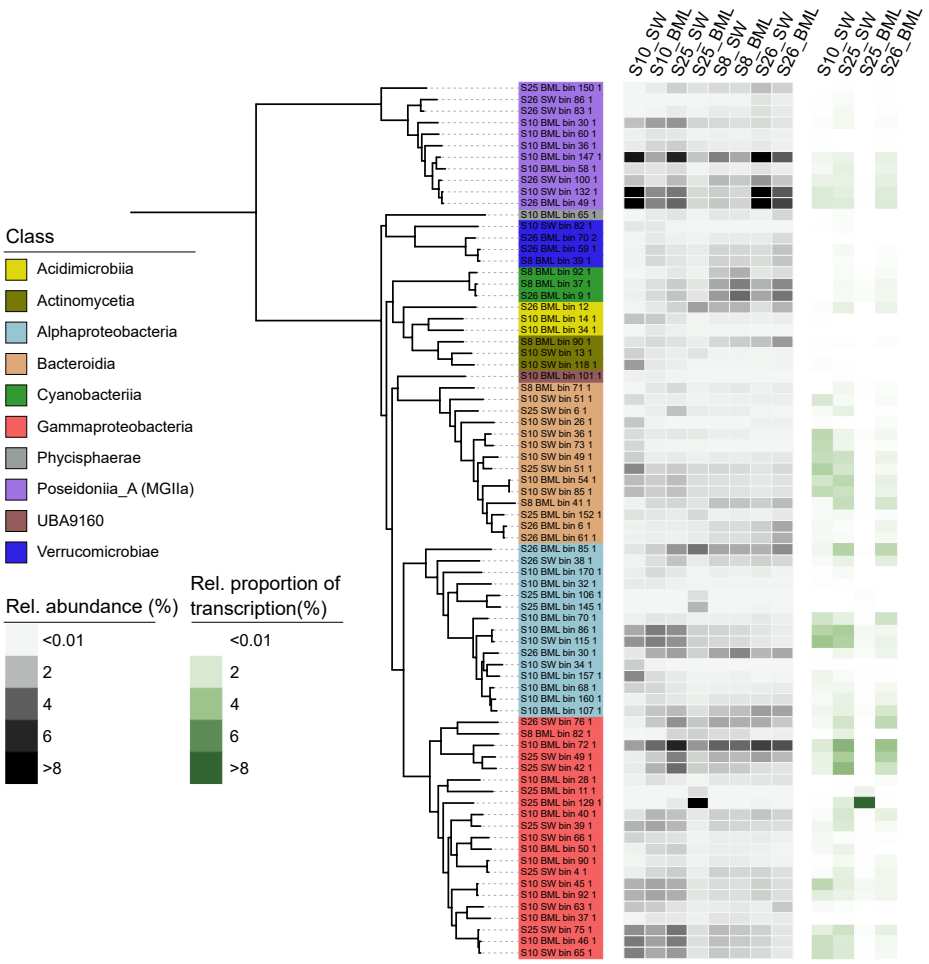


Figure 7. Population-representative metagenome-assembled genome phylogenetic tree with relative abundance and relative proportion of transcription. The tree was calculated from a concatenated alignment of 16 single-copy ribosomal protein genes (SC-RBPs), with a threshold of at least 8 genes per MAG for inclusion in the tree. Relative abundance of MAGs was defined as the quotient between the truncated average sequencing depth of the MAG and the average sequencing depth of 16 SC-RBPs in the sample. Relative proportion of transcription was defined by the average TPM value of the MAG-encoded 16 SC-RBPs and the whole sample average TPM values for the 16 SC-RBPs.

Carbohydrate utilisation potential was compared across the MAGs using CAZymes, TBDTs, sulfatases and peptidases. Peptidases were included for comparison, as proteins are another key substrate used by heterotrophic microbes, often simultaneously with carbohydrates. The largest CAZyme gene repertoires were observed in MAGs affiliated with *Verrucomicrobiae* and *Bacteroidia* that, on average, harboured 15 and 14 CAZymes per Mbp, respectively, whilst the lowest was observed in *Poseidoniia*-affiliated MAGs, 1 per Mbp (Figure 8). *Poseidoniia*-affiliated MAGs typically harboured high peptidase to CAZyme ratios, 3.7:1 on average, indicating a preference for proteinaceous substrates, in agreement with previous reports for members of this genera [42]. In addition to large CAZyme repertoires, *Verrucomicrobiae* MAGs also encoded the largest number of sulfatase genes, with 24 per Mbp, on average. Despite these broad taxa-related patterns, large variations were evident at the MAG level, with the number of CAZymes in *Bacteroidia* representatives ranging from 5 to 25 per Mbp – the

largest identified in an NS4-assigned MAG (S10_BML_bin_54_1) (Figure 8). The variations in genetic repertoires for carbohydrate degradation reflect previous descriptions of these taxa, where *Bacteroidia* typically harbour high numbers of CAZymes and *Verrucomicrobiae* are known to specialise in the degradation of fucose-containing sulfated polysaccharides [43, 44].

Comparing transcription levels of the focal gene groups across MAGs revealed distinct assemblages of populations that dominated carbohydrate utilisation in each sample (Figure 8), which reflected the taxonomic shifts observed at the community-level. However, the increased taxonomic resolution provides insights into the presence of single or multiple discrete populations assigned to each of the key, contributing genera. In sample S10_SW, CAZyme, sulfatase and TBDT transcription values were dominated by only a few *Bacteroidia* populations, which were taxonomically assigned to the *Formosa*, S25_SW_bin_51_1, followed by a MAG assigned to NS2b (SGZJ01), S10_SW_bin_49_1. Sample S25_SW was characterised by a larger number and diversity of populations of the *Bacteroidia* and *Gammaproteobacteria* that exhibited comparable transcription values. The key contributors to CAZyme transcription in S25_SW included two distinct populations affiliated with the NS4 (S10_SW_bin_85_1 and S10_BML_bin_54_1), *Formosa* (S25_SW_bin_51_1), SAR92

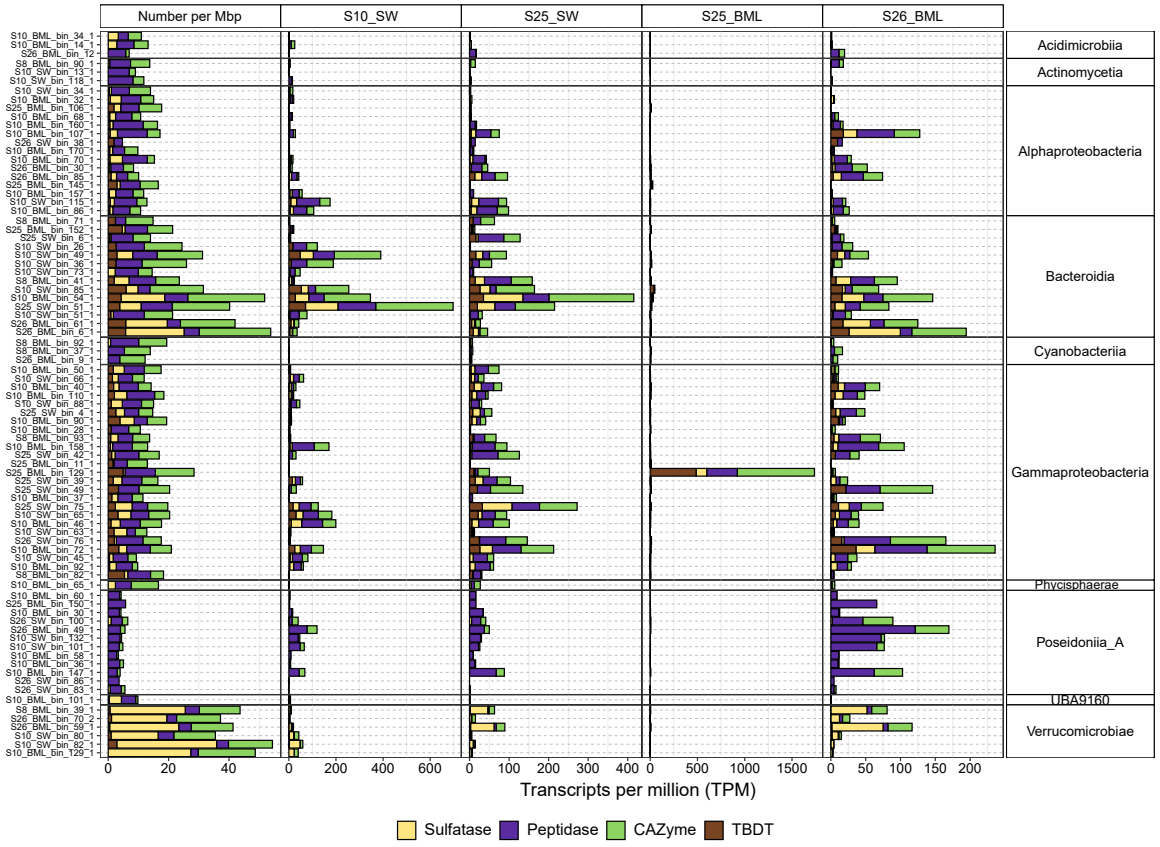


Figure 8. Count and transcription level of carbohydrate utilisation genes for population-representative MAGs. Carbohydrate utilisation genes were clustered into four groups based on function. The count and TPM value of all genes within each group were summed. Gene counts were further normalised by MAG genome size. TBDT = TonB-dependent transporter.

(S25_SW_bin_75_1) and SAR86 (S10_BML_bin_72_1). The highest TBDT transcription levels were also associated with these populations. High sulfatase transcription in S25_SW was observed in two populations of *Roseibacillus* (S8_BML_bin_39_1 and S26_BML_bin_59_1) that contributed little to transcription in S10_SW. In sample S26_BML, comparably high CAZyme gene transcription levels were observed across populations assigned to *Bacteroidia, Gammaproteobacteria, Poseidoniia, Alphaproteobacteria* and *Verrucomicrobiae.* In this sample, several distinct populations that contributed little in S10_SW and S25_SW, were amongst the largest contributors to CAZyme transcription, including those assigned to NS5 (S26_BML_bin_6_1), *Flavobacteriaceae* uncultured (S26_BML_bin_6_1) and *Planktomarina* (S10_BML_bin_107_1). As could be expected, the TPM values for the focal gene groups in S25_BML were dominated by the *Pseudoalteromonas* representative, with the large CAZyme TPM values being attributed to the CE11 gene identified at the community-level.

The MAG-based comparisons revealed highly dynamic population-level patterns in the key contributors to carbohydrate utilisation across the samples. This further illustrates the high spatial heterogeneity in carbohydrate utilisation and indicates that the taxonomic patterns observed at the community-level can be attributed to either single or multiple discrete populations.

To investigate the key substrates being targeted by these prominent populations and assess whether substrate-based niche partitioning played a role in the observed dynamics, we compared the composition of transcribed CAZyme gene families (Figure 9). From this, several notable patterns emerged. A higher diversity of transcribed CAZyme gene families was observed for *Bacteroidia* representatives compared to those assigned to *Gammaproteobacteria*, which reflects the evolution of the former as carbohydrate degrading specialists. Subsequently, the predicted substrate targets were also higher in number for *Bacteroidia* populations, although large differences were observed. The NS4 representative, S10_BML_bin_54_1, consistently exhibited the largest diversity of transcribed CAZyme gene families and, in some cases, dominated the transcription of these families at the community level; 90% of GH130 transcripts in S26_BML were attributed to this population. A similar observation was made for other CAZyme gene families, with >95% of GH125 transcripts attributed to a *Formosa* representative (S25_SW_bin_51_1) in three out of four samples. The predicted targets of these gene families include complex substrates, such as *α*-mannan (GH125) and *β*-mannan (GH130), which require specific genetic machinery for degradation that has been previously identified only in a minority of clades. In accordance with our findings, *Formosa* representatives were amongst the minority that harboured *α*-mannan degradation capacities in the microbial community responding to phytoplankton blooms [7]. This

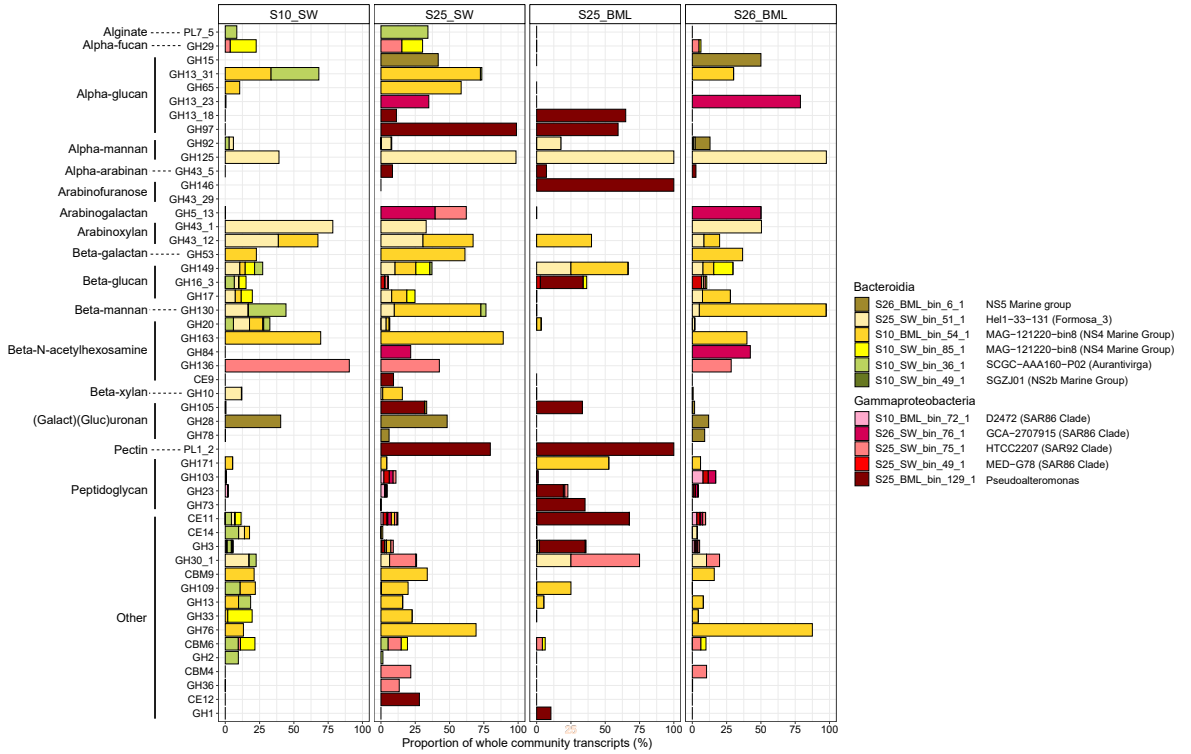


Figure 9. Comparison of CAZyme gene family transcription by selected MAGs across the four metatranscriptome samples. The selected MAGs included those that contributed >5% to total MAG carbohydrate-active enzyme (CAZyme) transcription values in each sample. To reduce complexity and focus on the key patterns, only the CAZyme families in each MAG whose transcription level was >=50% that of single-copy ribosomal protein gene transcription levels were included.

observation indicates substrate-based niche partitioning amongst the carbohydrate specialists within each sample. Furthermore, the profiles of transcribed CAZyme gene families by each specialist varied across samples, suggesting a shift in the substrates used (Figure 9).

The transcription of CAZyme gene families known to target labile, more simple polysaccharides, such as laminarin (GH16_3, GH17), were attributed to several populations in each sample, which together only represented a fraction of whole community transcription. Laminarin is a major storage compound of diatoms that consists of a linear β-1,3-linked glucose backbone with β-1,6-linked glucose branches. The production of laminarin by diatoms has been estimated at 12 ± 8 gigatons annually and can constitute 26 ± 17% of the particulate organic carbon pool [45]. The large production and relatively simple structure of laminarin makes it a prime target for heterotrophic bacteria. Amongst the *Bacteroidia* responders to phytoplankton blooms, the capacity to degrade laminarin was identified as a widespread trait [7, 11], which is in accordance with our observations. In contrast to the more specialised substrates described above, it is evident that some polysaccharides are more communal, providing accessible routes to carbon for numerous populations simultaneously.

Transcription of polysaccharide utilisation loci in MAGs

To further ameliorate distinctions in substrate utilisation patterns amongst the refined population selection, we compared the transcription levels of CAZymes found within polysaccharide utilisation loci (PULs) (Supplementary Figure S10). PULs are genetic regions which contain numerous genes involved in the degradation and uptake of specific polysaccharides. Here, we define PULs as all genetic loci that encode either a SusC/SusD pair along with two or more degradative CAZymes, three or more degradative CAZymes, or two degradative CAZymes along with several other genes involved in the transport and degradation of carbohydrates. Amongst the ten selected populations, six encoded PULs that were transcribed at different levels across the samples (Supplementary Figure S10 and S11). For example, the transcription of PUL_4 in the NS2b-affiliated MAG (S10_SW_bin_49_1), which is predicted to target glucomannan (GH3, GH16_3, GH130, GH92), was transcribed at equal levels to SC-RBP genes in sample S10_SW, but at much lower levels in the other three samples. Similarly, in the *Formosa*-assigned MAG (S25_SW_bin_51_1), the highest transcribed PUL in S10_SW was PUL_1, with α-mannan as a predicted target (GH92 x 2 and SusC/SusD pair), whilst PUL_5 was the most transcribed in S25_SW, with laminarin as a predicted target (GH149, GH17 x 3 and GH30_1). Such shifts in PUL transcription across samples, further supports the idea that the carbohydrate specialists, which house diverse CAZyme gene repertoires, are able to shift substrate target over spatial scales, likely in response to availability.

CONCLUSION

During late summer in the high North Atlantic, the carbohydrate fraction of POM and the utilisation of carbohydrates by microbial communities exhibits vertical and spatial heterogeneity. Monosaccharide and polysaccharide compositions of POM showed distinct differences across locations and depths. In particular, higher abundances were observed at locations above the continental slope compared to the open-ocean, which reflects the spatial differences and patchiness in primary productivity typically observed in this region in early summer [21]. Remnants of primary productivity from earlier seasonal stages, such as the presence of sulfated fucans that are known to accumulate in POM over weeks to months [28], consistently characterized the polysaccharide fraction in all samples. In addition, the varied detection and abundance of β-1,3-glucan, a simple and more labile substrate that persists for only days to a week in the POM pool [28], suggests local differences in recent primary production across the sampling locations. Concurrently, metatranscriptomics revealed the presence of active and dynamic populations, with distinct assemblages dominating carbohydrate utilisation in each sample and making use of labile (communal) and more complex (specialist) substrates. Carbohydrate specialists, such as those affiliated with *Formosa, Aurantivirga* and NS4 Marine Group, each exhibited distinct profiles of transcribed

CAZyme genes across samples, which indicated substrate-based niche partitioning. These specialists taxonomically resembled, at the genus-level, those known to be primary responders and the key carbohydrate degraders during phytoplankton blooms in temperate ecosystems, such as the North Sea. In combination, these results highlight that local biological and physical processes continue to shape the carbohydrate pool in a late summer stage and the key heterotrophic microbial populations continue to be responsive.

METHODS

Sample collection

Seawater samples were collected from ten stations located in the eastern Fram Strait and around the Svalbard archipelago in September 2020 during the MSM95 Maria S. Merian research cruise [46]. A map of the sampling locations (Figure 1) was generated using publically available bathymetric data [47, 48] and the QGIS v3.14.16-Pi [49] software. Seawater was collected using a Conductivity, Temperature and Depth (CTD)-rosette sampler from two distinct layers, including surface water (SW), typically 5 m depth, and the bottom of the surface mixed layer (BML), determined by sharp change in temperature, salinity and fluorescence values. One location was sampled twice over a two day period, with additional samples collected at 100 and 200 m depth, these are labelled as S1 and S6 (Supplementary Table S1). Of the water collected, 4 L was filtered sequentially through a 3 and 0.2 µm pore-size polycarbonate membrane filter (142 mm diameter) and immediately stored at -80 °C for ‘omics analysis. A second 4 L of seawater was filtered through a 0.7 µm pre-combusted GF/F filter (47 mm diameter) and immediately stored at -80 °C for monosaccharide and polysaccharide analysis.

Monosaccharide and polysaccharide analysis

The GF/F filters from all samples (ten stations and two depths) were cut into twelve equally-sized circular pieces, with a diameter of 11.2 mm. For monosaccharide analysis, two of the circular sections from each sample were subject to acid hydrolysis in glass ampules with 500 µl of 1 M HCl for 24 h at 100 °C. A 450 µl aliquot was then transferred to a speed vac for drying, before resuspension in 150 µl of MilliQ water. The resuspended material was briefly vortexed, spun down and transferred to a High-Performance Liquid Chromatography (HPLC) glass vial. The samples were analysed using High-Performance Anion Exchange Chromatography-Pulsed Amperometric Detection along with monosaccharide standards, as described in [28].

For polysaccharide analysis, the remaining ten circular sections from each sample were subject to a sequential extraction protocol with 1) MilliQ water, 2) 0.2 M EDTA (pH 7.5) and 3) 4 M NaOH with 0.1% w/v $NaBH_4$. In each round of extraction, filter sections were

placed in an Eppendorf tube with 450 µl of the respective solvent, briefly vortexed and then incubated for 2 h at 60 °C in a tube shaker at 650 rpm. After incubation, the tubes were centrifuged at 6000 x g for 15 min at 15 °C. The supernatant was transferred to a new tube and stored at -20 °C (except an aliquot that was directly used for microarray printing). The next solvent was added to the pellet and the same process was repeated.

The identification and semi-quantitative analysis of polysaccharide compounds was performed using a microarray-based approach, as described in [28]. Briefly, polysaccharide extracts along with twofold dilutions of each extract were loaded into wells of 384 microwell-plates and centrifuged at 3500 x *g* for 10 min at 15 °C. The contents of the wells were printed onto nitrocellulose membranes (0.45 µm pore size) using a microarray robot (Sprint, Arrayjet, Roslin, UK) under conditions of 20 °C and 50% humidity. Extracts were printed in duplicates, so that each extract was represented by 4 spots in the microarray (2 x extracts and 2 x diluted extracts). The printed arrays were blocked for 1 h in 1 x PBS with 5% (w/v) non-fat milk powder (MPBS). The MPBS was subsequently discarded and the microarrays were individually incubated for 2 h with one of 50 polysaccharide-specific monoclonal antibodies (Supplementary Table S11). After incubation, the arrays were washed in PBS and incubated for 2 h in anti-rat, anti-mouse or anti-His tag secondary antibodies conjugated to alkaline phosphatase diluted 1:5000, 1:5000 and 1:1500, respectively. Arrays were thoroughly washed in PBS, followed by deionised water and developed in a solution containing 5-bromo-4-chloro-3-indolylphosphate and nitro blue tetrazolium in alkaline phosphatase buffer (100 mM NaCl, 5 mM $MgCl_2$, 100 mM Tris-HCl, pH 9.5). Developed arrays were scanned and the binding of each probe against each spotted sample was quantified using Array-Pro Analyzer 6.3 (Media Cybernetics). Mean antibody signal binding intensity was then determined for each extract. The highest mean value of the dataset was set to 100 and all other values were normalised accordingly.

Metagenome and metatranscriptome sequencing

Filtered seawater samples of the 0.2 – 3 µm fraction from SW and BML depths of four different stations were subject to a dual nucleic acid isolation protocol using the DNA/RNA Mini Prep Plus kit from Zymo Research (Irvine, CA, USA), according to the manufacturer's instructions. The quality of extracted DNA was assessed using capillary electrophoresis with a FEMTOpulse (Agilent), whilst RNA quality was assessed using a PicoChip on a Bioanalyser (Agilent, CA, USA). Ultra-low DNA libraries were prepared from the eight samples without further fragmentation by the protocol "Procedure & Checklist - Preparing HiFi SMRTbell® Libraries from Ultra-Low DNA Input" of PacBio (CA, USA). Libraries were inspected again on FEMTOpulse and sequenced on 4 x 8M SMRT cells on a Sequel II platform for 30 h with sequencing chemistry 2.0 and binding kit 2.0 (two samples multiplexed per SMRT cell). Four

of the samples were additionally selected for metatranscriptome sequencing (S10_surface, S25_Surface, S25_BML, S26_BML). The extracted RNA from the four was quality assessed with a PicoChip on a Bioanalyser (Agilent). Illumina-compatible libraries were produced using the Universal Prokaryotic RNA-Seq library preparation kit, incl. Prokaryotic AnyDeplete® (Tecan Genomics, CA, USA). Libraries were sequenced on a HiSeq 3000 platform with 2 x 150 bp paired-end read mode.

HiFi read taxonomic classification

To taxonomically classify the HiFi reads, we used a custom pipeline with a GTDB-based protein database. To generate the database, gene amino acid sequences of all GTDB species-representatives (release 207) were obtained from https://data.ace.uq.edu.au/public/gtdb/data/releases/release207/207.0/ and subsequently clustered at 99% sequence identity, to remove redundancy. The representative gene sequences were then used to generate NCBI-style taxdump files (nodes.dmp, names.dmp and accession2taxid) using scripts from https://github.com/nick-youngblut/, and converted to a Diamond blast [50] database.

Open reading frames were predicted on raw HiFi reads using prodigal v2.6.3. The reads were then aligned against the generated GTDB-based protein database using Diamond blastp (v0.9.14; parameters: --id 50 --top 5 --fast). A secondary filtering step was applied to the output including identity threshold >65%, e-value <1E-10 and query-cover >50%. Using the remaining hits, a single taxonomic classification for each gene was determined using a last common ancestor approach, *lca* command from TaxonKit [51]. To further increase the number of genes taxonomically classified, the last common ancestor algorithm was then applied to all genes within a single HiFi read, resulting in a single taxonomic classification for each HiFi read, and it's containing genes.

HiFi read functional annotation

For functional characterization, HiFi reads were subject to a custom annotation pipeline, similar to that described in Priest *et al.* [52], with some additions. In brief, open reading frames were predicted and genes annotated using Prokka v1.14.6 [53]. Transporter genes were annotated via HMMscan against the tcDoms profiles of the Transporter Classification Database (downloaded November 2021), using the family-specific gathering cut-off threshold (parameters: --cut_ga). The presence of CAZymes was determined by a dual annotation using HMMscan against the dbCAN v10 database [54] and Diamond blastp search against the CAZy database (release 09242021) [55]. Sulfatases were annotated via HMMscan against Pfam sulfatase family PF0084 and blastp search against the SulfAtlas v1.3 [56] database. Peptidases were annotated using a blastp search against the MEROPS database [57]. TonB-

dependent receptors were predicted by HMMscan against 15 different TIGRFAM profiles (detailed information provided in Supplementary Material S1). Lastly, SusD genes were predicted via HMMscan against the Pfam profiles PF12741, PF12771, PF14322, PF07980. All annotations were combined into a single table for each metagenome.

Single-copy ribosomal protein (SC-RBP) gene analysis

Based on the annotations from the above-described pipeline, 16 single-copy ribosomal protein (SC-RBP) genes were extracted from each metagenome. The average number of SC-RBP was used as a proxy for the number of genomes recovered in each metagenome. A subset of four SC-RBP genes (RBP L3, L4, L6 and S8) were further used to determine the number of species sampled in each metagenome. Sequences belonging to this subset from each sample were clustered at previously defined gene-specific ANI thresholds [31] and the average number of clusters generated from the four, was defined as the number of species in each sample. In order to characterize and compare community composition across samples, the RBP L6 was selected, based on its high recoverability and ANI species delineation accuracy [31]. Taxonomic classification of RBP L6 gene sequences was derived from the read taxonomy.

Assembly, binning and metagenome-assembled genome recovery

The assembly of Hifi reads was performed using MetaFlye v2.8 [58] (parameters: --meta –pacbio-hifi –hifi-error 0.01 –keep-haplotypes) with a slight modification. The alignment.py script supplied by MetaFlye was modified to call the ‘map-hifi’ preset with Minimap2 instead of the ‘map-pb’ preset. The Minimap2 tool [59] is used for read alignment to the assembly graph and subsequent contig sequence generation. By modifying the preset specified, the % identity of the read mapping is increased. To generate contig coverage information, raw reads were mapped to contigs using Minimap2 v2.1 with the ‘map-hifi’ preset. Contigs were binned using Metabat2 [60], Concoct v1.1 [61] and Vamb v3.0.2 [62], with the output from all tools being provided to DasTool v1.1.2 [63], which performs a scoring and contig dereplication strategy to generate a single consensus set of bins. The consensus set of bins were subject to manual refinement using the Anvi’o v7 [64] interactive interface to generate MAGs. MAGs were dereplicated at a 98% ANI threshold using dRep v3.2.2 (parameters: --comp 50 --con 5 --sa 0.98 --nc 0.6). The completeness and contamination of representative MAGs was estimated using CheckM v1.1.2 [65].

The taxonomic classification of MAGs was performed using two methods. Firstly, the classify_wf pipeline of GTDB-tk v1.0.2 [66] (Release 207) was used. Secondly, where possible, 16S rRNA gene sequences were extracted from MAGs using Barrnap v0.9 [67] and taxonomically classified following the same process described above in ‘*Phylogenetic*

characterisation of communities'. Of the species-representative MAGs, 84% contained a complete 16S rRNA gene and were assigned a taxonomy from the SILVA database.

MAG relative abundance estimation

The relative abundance of representative MAGs was determined using a similar approach to Orellana et al. [44]. In brief, reads were competitively recruited from each metagenome to the MAG representatives. Mapped reads were converted into depth values using Genomecov (-bga option) from the Bedtools package [68] and the 80% central truncated average of the sequencing depth (TAD) was determined using the 'BedGraph.tad.rb' script (option range 80) from the enveomics collection [69]. The relative abundance was then determined as the quotient between the TAD value and the average sequencing depth of the 16 SC-RBPs of each MAG.

MAG functional characterisation

The functional characteristaion of MAGs was performed following the same procedure described for the raw HiFi reads above except for an additional process of polysaccharide utilisation loci (PULs) detection. PULs were defined as genetic loci containing a SusC/SusD gene pair with two or more degradative CAZymes or the presence of at least three degradative CAZymes in close proximity (maximum six genes apart). PULs were manually inspected and visualised at BioRender.com.

Transcription level of genes at the community- and MAG-level

Adapters and low quality reads were removed from the metatranscriptomes using BBDuk of the BBtools program v38.73 (http://bbtools.jgi.doe.gov) (parameters: ktrim=r, k=29, mink=12, hdist=1, tbo=t, tpe=t, qtrim=rl, trimq=20, minlength=100). Although an rRNA depletion step was performed prior to sequencing library construction, it is expected that 5 – 15% of reads would still be related to rRNA. As such, SortMeRNA v2.0 [71] was used to filter out rRNA sequences from the dataset, with the SILVA SSU Ref 138 NR99 database as a reference. The transcription level of genes was determined by read recruitment of transcripts to the predicted gene sequences from the raw HiFi reads and the MAGs using BBmap of the BBtools program (98% identity threshold). Mapped read values were converted to TPM, according to Wagner *et al.* [72], using the number of transcripts mapped to the raw HiFi read-predicted genes as the total transcript values. In order to compare the transcription level of MAGs across samples, we determined the average TPM value of the 16 SC-RBPs for each MAG in each sample, and took the quotient of this and the average TPM value of the 16 SC-RBPs in the whole sample – providing proportional transcription of all genomes recovered.

DATA AVAILABILITY

The measurements of several abiotic parameters from sensors mounted on the CTD have been published under the PANGAEA accession 943220 [73]. The monosaccharide concentrations have been deposited under the PANGAEA accession xxx (Not yet available). The metagenomic raw reads, assemblies and metagenome-assembled genomes along with the metatranscriptomic raw reads were deposited at ENI-EBA, under the project accession xxx (Not yet available).

AUTHOR CONTRIBUTIONS

TP wrote the manuscript, performed the metagenomic, metatranscriptomic and carbohydrate data analysis. TP and SVM extracted the carbohydrates and SVM subsequently carried out the microarray analysis. SVM, JHH, BMF and RA contributed to the interpretation of the results and the formulation of the story. BMF and TP planned the work and devised the project. All authors contributed to reviewing and improving the manuscript.

ACKNOWLEDGEMENTS

We would like to thank the captain and crew of the *RV. Maria. S. Merian* for their support throughout all sampling aspects of this project. We thank Alek Bolte for his assistance with HPAEC-PAD. We thank Murat Eren for his invaluable input and ideas with respect to data analysis.

REFERENCES

1. Benner R, Pakulski JD, Mccarthy M, Hedges JI, Hatcher PG. Bulk chemical characteristics of dissolved organic matter in the ocean. *Science*. 1992;255:1561–1564.

2. Myklestad S. Production of carbohydrates by marine planktonic diatoms. I. Comparison of nine different species in culture. *J Exp Mar Biol Ecol*. 1974;15:261–274.

3. Granum E, Kirkvold S, Myklestad SM. Cellular and extracellular production of carbohydrates and amino acids by the marine diatom Skeletonema costatum: diel variations and effects of N depletion. *Mar Ecol Prog Ser*. 2002;242:83–94.

4. Bellinger B, Abdullahi A, Gretz M, Underwood G. Biofilm polymers: relationship between carbohydrate biopolymers from estuarine mudflats and unialgal cultures of benthic diatoms. *Aquat Microb Ecol*. 2005;38:169–180.

5. Abdullahi AS, Underwood GJC, Gretz MR. Extracellular matrix assembly in diatoms (Bacillariophyceae). V. Environmental effects on polysaccharide synthesis in the model diatom, Phaeodactylum Tricornutum1. *J Phycol*. 2006;42:363–378.

6. Reintjes G, Arnosti C, Fuchs BM, Amann R. An alternative polysaccharide uptake mechanism of marine bacteria. *ISME J*. 2017;11:1640–1650.

7. Krüger K, Chafee M, Ben Francis T, Glavina del Rio T, Becher D, Schweder T, et al. In marine Bacteroidetes the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. *ISME J*. 2019;13:2800–2816.

8. Arnosti C, Wietz M, Brinkhoff T, Hehemann J-H, Probandt D, Zeugner L, et al. The Biogeochemistry of marine polysaccharides: sources, inventories, and bacterial drivers of the carbohydrate cycle. *Annu Rev Mar Sci*. 2021;13:81–108.

9. Henrissat B. A classification of glycosyl hydrolases based on amino acid sequence similarities. *Biochem J*. 1991;280:309–316.

10. 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.

11. Kappelmann L, Krüger K, Hehemann J-H, Harder J, Markert S, Unfried F, et al. Polysaccharide utilization loci of North Sea Flavobacteriia as basis for using SusC/D-protein expression for predicting major phytoplankton glycans. *ISME J*. 2019;13:76–91.

12. Hoarfrost A, Arnosti C. Heterotrophic extracellular enzymatic activities in the Atlantic Ocean follow patterns across spatial and depth regimes. *Front Mar Sci*. 2017;4.

13. D'Ambrosio L, Ziervogel K, MacGregor B, Teske A, Arnosti C. Composition and enzymatic function of particle-associated and free-living bacteria: a coastal/offshore comparison. *ISME J*. 2014;8:2167–2179.

14. Arnosti C, Steen AD, Ziervogel K, Ghobrial S, Jeffrey WH. Latitudinal gradients in degradation of marine dissolved organic carbon. *PLOS ONE*. 2011;6:e28900.

15. 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*. 2012;336:608–611.

16. Teeling H, Fuchs BM, Bennke CM, Krüger K, Chafee M, Kappelmann L, et al. Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife*. 2016;5:e11888.

17. Avcı B, Krüger K, Fuchs BM, Teeling H, Amann RI. Polysaccharide niche partitioning of distinct Polaribacter clades during North Sea spring algal blooms. *ISME J*. 2020;14:1369–1383.

18. Nilsen F, Cottier F, Skogseth R, Mattsson S. Fjord–shelf exchanges controlled by ice and brine production: The interannual variation of Atlantic Water in Isfjorden, Svalbard. *Cont Shelf Res*. 2008;28:1838–1853.

19. Richter ME, von Appen W-J, Wekerle C. Does the East Greenland Current exist in the northern Fram Strait? *Ocean Sci*. 2018;14:1147–1165.

20. Skoog A, Benner R. Aldoses in various size fractions of marine organic matter: Implications for carbon cycling. *Limnol Oceanogr*. 1997;42:1803–1813.

21. von Jackowski A, Grosse J, Nöthig E-M, Engel A. Dynamics of organic matter and bacterial activity in the Fram Strait during summer and autumn. *Philos Trans R Soc Math Phys Eng Sci*. 2020;378:20190366.

22. Tanoue E, Handa N. Monosaccharide composition of marine particles and sediments from the Bering Sea and northern North Pacific. *Oceanol Acta*. 1987;10:91–99.

23. Liebezeit G, Bölter M. Water-extractable carbohydrates in particulate matter of the Bransfield Strait. *Mar Chem*. 1991;35:389–398.

24. Compiano A-M, Romano J-C, Garabetian F, Laborde P, de la Giraudièrea I. Monosaccharide composition of particulate hydrolysable sugar fraction in surface microlayers from brackish and marine waters. *Mar Chem*. 1993;42:237–251.

25. Ittekkot V, Brockmann U, Michaelis W, Degens E. Dissolved Free and Combined Carbohydrates During a Phytoplankton Bloom in the Northern North Sea. *Mar Ecol Prog Ser*. 1981;4:299–305.

26. Madadi R, Maljaee H, Serafim LS, Ventura SPM. Microalgae as contributors to produce biopolymers. *Mar Drugs*. 2021;19:466.

27. Huang G, Vidal-Melgosa S, Sichert A, Becker S, Fang Y, Niggemann J, et al. Secretion of sulfated fucans by diatoms may contribute to marine aggregate formation. *Limnol Oceanogr*. 2021;66:3768–3782.

28. Vidal-Melgosa S, Sichert A, Francis TB, Bartosik D, Niggemann J, Wichels A, et al. Diatom fucan polysaccharide precipitates carbon during algal blooms. *Nat Commun*. 2021;12:1150.

29. Boysen AK, Carlson LT, Durham BP, Groussman RD, Aylward FO, Ribalet F, et al. Particulate metabolites and transcripts reflect diel oscillations of microbial activity in the surface ocean. *mSystems*. 2021;6:e00896-20.

30. van Oijen T, van Leeuwe MA, Gieskes WWC. Variation of particulate carbohydrate pools over time and depth in a diatom-dominated plankton community at the Antarctic Polar Front. *Polar Biol*. 2003;26:195–201.

31. Olm MR, Crits-Christoph A, Diamond S, Lavy A, Matheus Carnevali PB, Banfield JF. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems*. 2020;5:e00731-19.

32. Wietz M, Bienhold C, Metfies K, Torres-Valdés S, von Appen W-J, Salter I, et al. The polar night shift: seasonal dynamics and drivers of Arctic Ocean microbiomes revealed by autonomous sampling. *ISME Commun*. 2021;1:1–12.

33. Fadeev E, Salter I, Schourup-Kristensen V, Nöthig E-M, Metfies K, Engel A, et al. Microbial communities in the east and west Fram Strait during sea ice melting season. *Front Mar Sci*. 2018;5.

34. von Jackowski A, Becker KW, Wietz M, Bienhold C, Zäncker B, Nöthig E-M, et al. Variations of microbial communities and substrate regimes in the eastern Fram Strait between summer and fall. *Environ Microbiol*. 2022;24:4124–4136.

35. Priest T, Heins A, Harder J, Amann R, Fuchs BM. Niche partitioning of the ubiquitous and ecologically relevant NS5 marine group. *ISME J*. 2022;16:1570–1582.

36. Rinta-Kanto JM, Sun S, Sharma S, Kiene RP, Moran MA. Bacterial community transcription patterns during a marine phytoplankton bloom. *Environ Microbiol*. 2012;14:228–239.

37. Pontiller B, Martínez-García S, Joglar V, Amnebrink D, Pérez-Martínez C, González JM, et al. Rapid bacterioplankton transcription cascades regulate organic matter utilization during phytoplankton bloom progression in a coastal upwelling system. *ISME J*. 2022;1–13.

38. Holmström C, Kjelleberg S. Marine Pseudoalteromonas species are associated with higher organisms and produce biologically active extracellular agents. *FEMS Microbiol Ecol*. 1999;30:285–293.

39. Moisander PH, Sexton AD, Daley MC. Stable associations masked by temporal variability in the marine copepod microbiome. *PLOS ONE*. 2015;10:e0138967.

40. Shoemaker KM, Moisander PH. Microbial diversity associated with copepods in the North Atlantic subtropical gyre. *FEMS Microbiol Ecol*. 2015;91:fiv064.

41. Francis B, Urich T, Mikolasch A, Teeling H, Amann R. North Sea spring bloom-associated Gammaproteobacteria fill diverse heterotrophic niches. *Environ Microbiome*. 2021;16:15.

42. Orellana LH, Ben Francis T, Krüger K, Teeling H, Müller M-C, Fuchs BM, et al. Niche differentiation among annually recurrent coastal Marine Group II Euryarchaeota. *ISME J*. 2019;13:3024–3036.

43. 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.

44. Orellana LH, Francis TB, Ferraro M, Hehemann J-H, Fuchs BM, Amann RI. Verrucomicrobiota are specialist consumers of sulfated methyl pentoses during diatom blooms. *ISME J*. 2021;1–12.

45. Becker S, Tebben J, Coffinet S, Wiltshire K, Iversen MH, Harder T, et al. Laminarin is a major molecule in the marine carbon cycle. *Proc Natl Acad Sci*. 2020;117:6599–6607.

46. Purser A, Hoge U, Busack M, Hagemann J, Lehmenhecker S, Dauer E, et al. Arctic Seafloor Integrity, Cruise No. MSM95 - (GPF 19-2_05), 09.09.2020 - 07.10.2020, Emden (Germany). 2020.

47. GEBCO Compilation Group. GEBCO 2020 Grid. 2020.

48. Jakobsson M, Mayer LA, Bringensparr C, Castro CF, Mohammad R, Johnson P, et al. The international bathymetric chart of the Arctic Ocean version 4.0. *Sci Data*. 2020;7:176.

49. QGIS.org. 2021. QGIS Association.

50. Buchfink B, Xie C, Huson DH. Fast and sensitive protein alignment using DIAMOND. *Nat Methods*. 2015;12:59–60.

51. Shen W, Ren H. TaxonKit: A practical and efficient NCBI taxonomy toolkit. *J Genet Genomics*. 2021;48:844–850.

52. Priest T, Orellana LH, Huettel B, Fuchs BM, Amann R. Microbial metagenome-assembled genomes of the Fram Strait from short and long read sequencing platforms. *PeerJ*. 2021;9:e11721.

53. Seemann T. Prokka: rapid prokaryotic genome annotation. *Bioinforma Oxf Engl*. 2014;30:2068–2069.

54. Zhang H, Yohe T, Huang L, Entwistle S, Wu P, Yang Z, et al. dbCAN2: a meta server for automated carbohydrate-active enzyme annotation. *Nucleic Acids Res*. 2018;46:95–101.

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

56. Barbeyron T, Brillet-Guéguen L, Carré W, Carrière C, Caron C, Czjzek M, et al. Matching the diversity of sulfated biomolecules: creation of a classification database for sulfatases reflecting their substrate specificity. *PLOS ONE*. 2016;11:e0164846.

57. Rawlings ND, Barrett AJ, Thomas PD, Huang X, Bateman A, Finn RD. The MEROPS database of proteolytic enzymes, their substrates and inhibitors in 2017 and a comparison with peptidases in the PANTHER database. *Nucleic Acids Res*. 2018;46:624–632.

58. Kolmogorov M, Bickhart DM, Behsaz B, Gurevich A, Rayko M, Shin SB, et al. metaFlye: scalable long-read metagenome assembly using repeat graphs. *Nat Methods*. 2020;17:1103–1110.

59. Li H. Minimap2: pairwise alignment for nucleotide sequences. *Bioinformatics*. 2018;34:3094–3100.

60. Kang DD, Li F, Kirton E, Thomas A, Egan R, An H, et al. MetaBAT 2: an adaptive binning algorithm for robust and efficient genome reconstruction from metagenome assemblies. *PeerJ*. 2019;7:e7359.

61. Alneberg J, Bjarnason BS, de Bruijn I, Schirmer M, Quick J, Ijaz UZ, et al. Binning metagenomic contigs by coverage and composition. *Nat Methods*. 2014;11:1144–1146.

62. Nissen JN, Johansen J, Allesøe RL, Sønderby CK, Armenteros JJA, Grønbech CH, et al. Improved metagenome binning and assembly using deep variational autoencoders. *Nat Biotechnol*. 2021;39:555–560.

63. Sieber CMK, Probst AJ, Sharrar A, Thomas BC, Hess M, Tringe SG, et al. Recovery of genomes from metagenomes via a dereplication, aggregation and scoring strategy. *Nat Microbiol*. 2018;3:836–843.

64. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*. 2015;3:e1319.

65. 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.

66. Chaumeil P-A, Mussig AJ, Hugenholtz P, Parks DH. GTDB-Tk: a toolkit to classify genomes with the Genome Taxonomy Database. *J Bioinform*. 2020;36:1925–1927.

67. Seeman T. barrnap 0.9: rapid ribosomal RNA prediction.

68. Quinlan AR, Hall IM. BEDTools: a flexible suite of utilities for comparing genomic features. *J Bioinform*. 2010;26:841–842.

69. Rodriguez-R LM, Konstantinidis KT. The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Prepr*. 2016;4:e1900v1.

70. Bushnell B. BBTools software package. http://bbtools.jgi.doe.gov. .

71. Kopylova E, Noé L, Touzet H. SortMeRNA: fast and accurate filtering of ribosomal RNAs in metatranscriptomic data. *J Bioinform*. 2012;28:3211–3217.

72. Wagner GP, Kin K, Lynch VJ. Measurement of mRNA abundance using RNA-seq data: RPKM measure is inconsistent among samples. *Theory Biosci*. 2012;131:281–285.

73. Priest T, Merten V, Purser A. Physical oceanography during RV MARIA S. MERIAN cruise MSM95. 2022. PANGAEA.

Supplementary Information S1

SUPPLEMENTARY RESULTS

Comparisons of microbial community functionality

Prior to assessing carbohydrate utilisation patterns, we first assessed differences in whole community functionality and how this compared to differences in phylogenetic composition. Across the eight metagenomes, a total of ~25 million genes were identified, which corresponded to 17,546 unique functional annotations – this equated to a functional annotation rate of ~50%. Furthermore, we determined TPM values for these genes in each metatranscriptome. To more reliably compare changes across samples, gene counts were normalized by the number of genomes predicted in each metagenome, hereon referred to as normalized gene count (NGC). Employing a dissimilarity-based approach resulted in distinct patterns in community functionality that did not clearly correspond to shifts in phylogeny (Supplementary Figure S7). Sample S10_SW was the most dissimilar to all other samples based on phylogeny, whilst S25_BML was the most dissimilar based on functional gene and transcribed functional gene composition. In addition, higher similarity was observed within stations for S8 and S26 samples based on phylogeny, whilst only S8 samples clustered together based on functional gene composition. These patterns indicate a degree of decoupling between phylogeny and function in the samples but a degree of coherency in the functional potential and the actively transcribed functions.

Metagenome-assembled genomes and population-specific dynamics

To provide higher resolution insights into spatial heterogeneity in carbohydrate utilisation by microbial communities, we recovered and analysed metagenome-assembled genomes (MAGs). We recovered 83 population-representative MAGs (99% ANI delineation threshold), 17 of which are classified as high-quality drafts and 66 as medium-quality drafts according to MIMAGs standards [23]. Genome sizes ranged from 0.863 – 5.1 Mbp, with an average of 2.1 Mbp, and the number of contigs ranged from 3 – 117, with an average of 37 (Supplementary Table S1). In addition, 85% of the MAGs contained at least one complete rRNA operon, allowing for a dual taxonomic classification against the SILVA and GTDB databases. The phylogenetic diversity captured by the MAGs included 10 classes and 61 genera, with the most MAG-rich classes being *Alphaproteobacteria* (n=15), *Bacteroidia* (n=14), *Gammaproteobacteria* (n=25) and *Poseidoniia_A* (n=12) (Supplementary Table S1).

The recovered representative MAGs captured between 48 – 88% of the metagenomic reads and 8 – 23% of the metranscriptomic reads. Across the metagenomes, MAG-specific dynamics were evident, which included sample-, depth- and station-specific patterns (Figure

7). The recovery of a MAG classified as *Pseudoalteromonas primoryensis,* which constituted 22.8% relative abundance of sample S25_BML, further indicates a single species-bloom that was captured in this sample. Aside from this, the MAGs reaching the highest relative abundances were associated with the MGIIa-L1 genus, including S26_BML_bin_49_1 (12.3%), S10_SW_bin_101_1 (12.2%) and S10_SW_bin_132_1 (10.2%) (Figure 8). Similar to the observations in community-level analysis, higher relative abundance of a MAG did not always correspond to higher proportion of transcription.

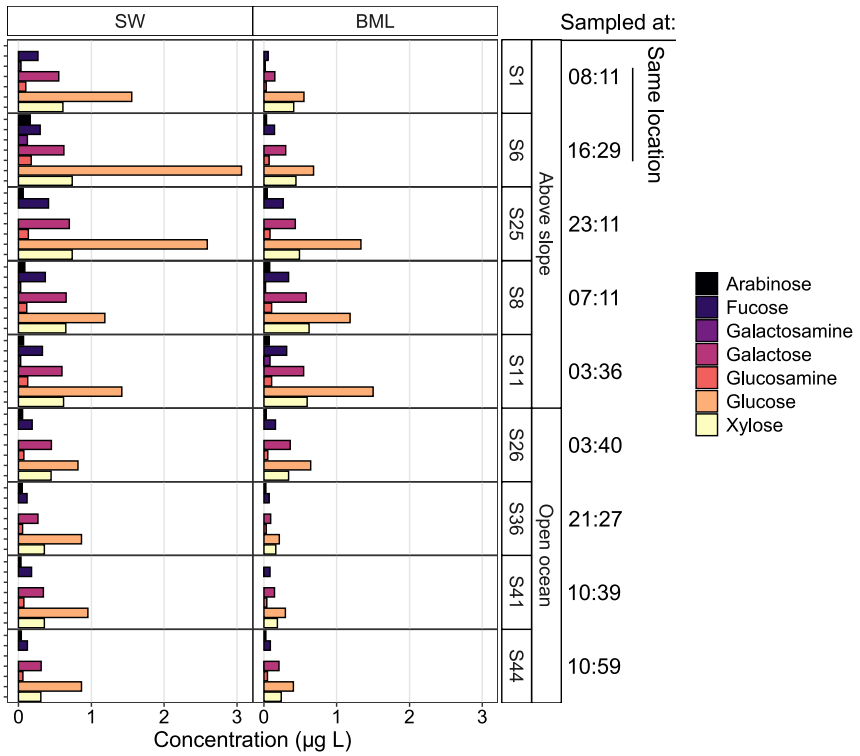
Transcription of polysaccharide utilisation loci in MAGs

To further ameliorate distinctions in substrate utilisation patterns amongst the refined MAG selection, we compared the transcription levels of CAZymes found within polysaccharide utilisation loci (PULs) (Figure 10). PULs are genetic regions which contain numerous genes involved in the degradation and uptake of specific polysaccharides. Here, we define PULs as all genetic loci that encode either a SusC/SusD pair along with two or more degradative CAZymes, three or more degradative CAZymes or two degradative CAZymes along with several other genes involved in the transport and degradation of carbohydrates. Amongst the ten MAGs that contributed >5% of total MAG CAZyme TPM in each sample, six encoded PULs that exhibited different transcription levels in relation to SC-RBP transcription across the samples (Supplementary Figure S9 and S10). For example, the CAZymes located on PUL_4 in the NS2b Marine Group-affiliated MAG (S10_SW_bin_49_1), which is predicted to target glucomannan (GH3, GH16_3, GH130, GH92), were transcribed at equal levels to housekeeping genes in sample S10_SW, but at much lower levels in the other three samples. Similarly, in the *Formosa*-assigned MAG (S25_SW_bin_51_1), the highest transcribed PUL in S10_SW was PUL_1, with α-mannan as a predicted target (GH92 x 2 and SusC/SusD pair), whilst PUL_5 was the most transcribed in S25_SW, with laminarin as a predicted target (GH149, GH17 x 3 and GH30_1).

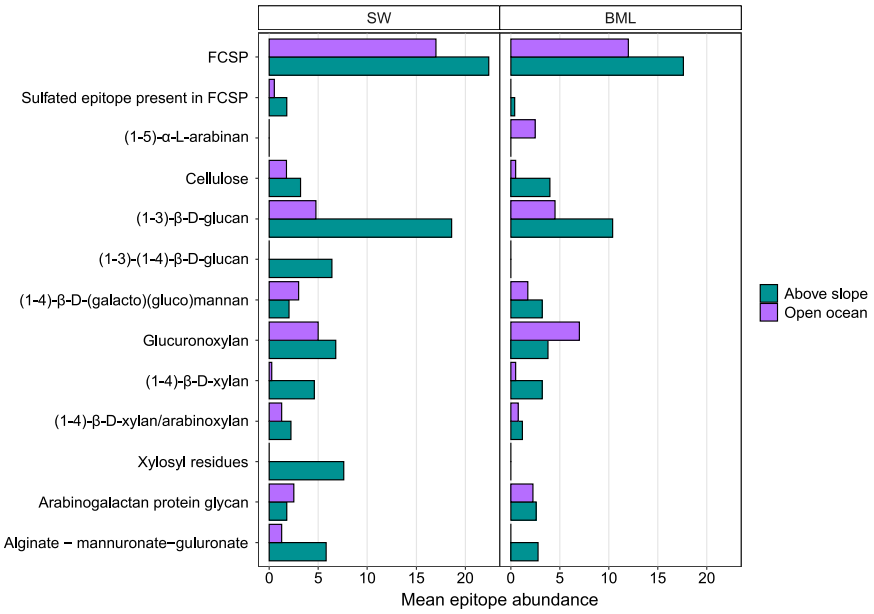
Supplementary tables and figures

All supplementary tables are available on the USB attached with this thesis and at the following link. Due to their large size, they were not included in the printed thesis.

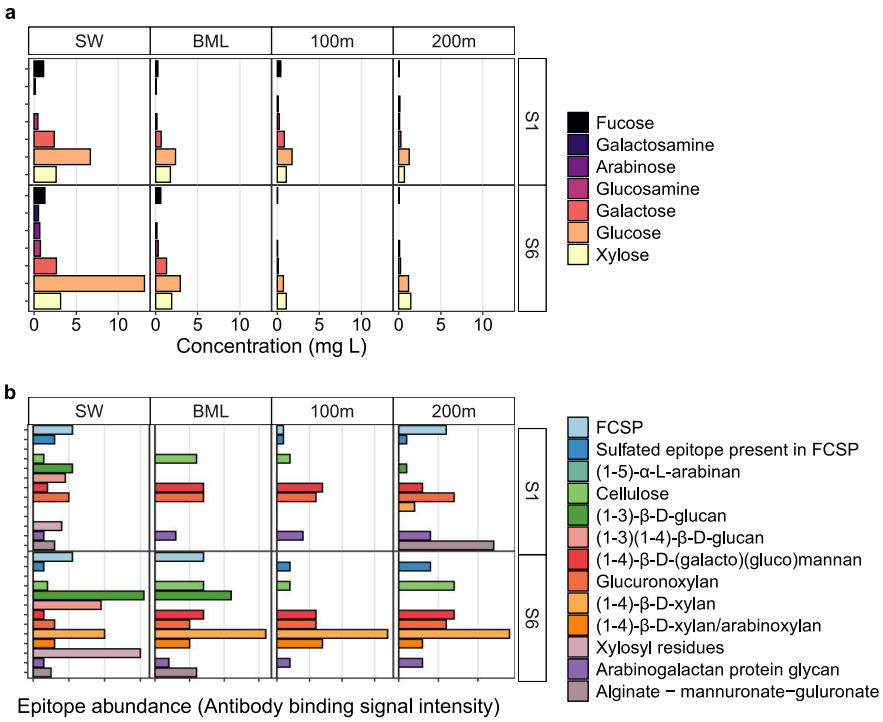
https://owncloud.mpi-bremen.de/index.php/s/mcU1bXHvHdtRsAB



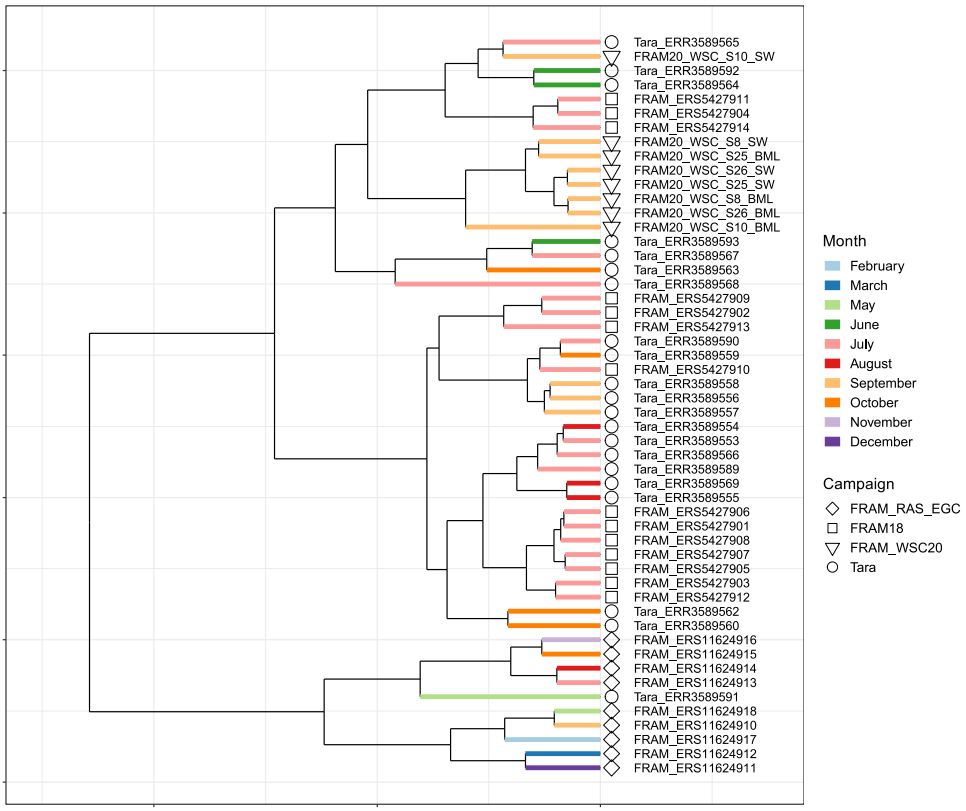
Supplementary Figure S1. Monosaccharide concentrations of particulate organic matter across surface water (SW) and bottom of mixed layer (BML) samples.



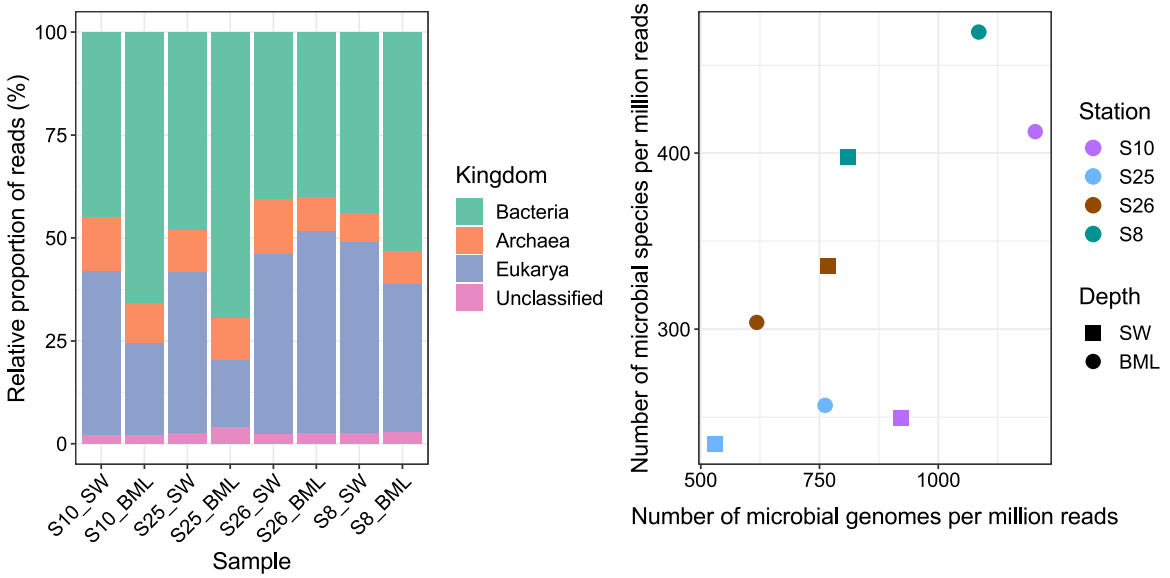
Supplementary Figure S2. The mean abundance of polysaccharide epitopes in particulate organic matter in above slope compared to open ocean samples. SW = surface water, BML = bottom of surface mixed layer. Mean abundances were determined by taking the average values from the respective stations. FCSP = fucose-containing sulphated polysaccharide.



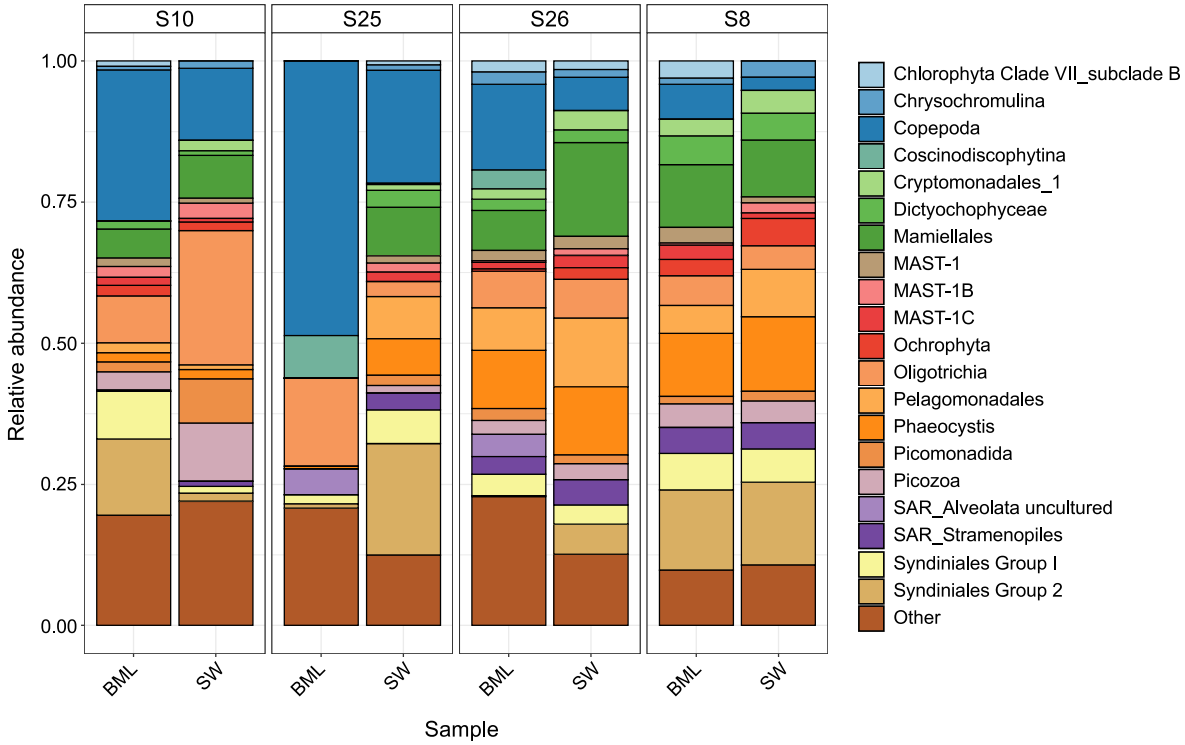
Supplementary Figure S3. Monosaccharide and polysaccharide compositions in particulate organic matter from the same sampling location over a two day period. Station S1 samples were collected at 08:11 whilst S6 samples were collected at 16:30 (+24 h). FCSP = fucose-containing sulphated polysaccharide, SW = surface water, BML = bottom of surface mixed layer.



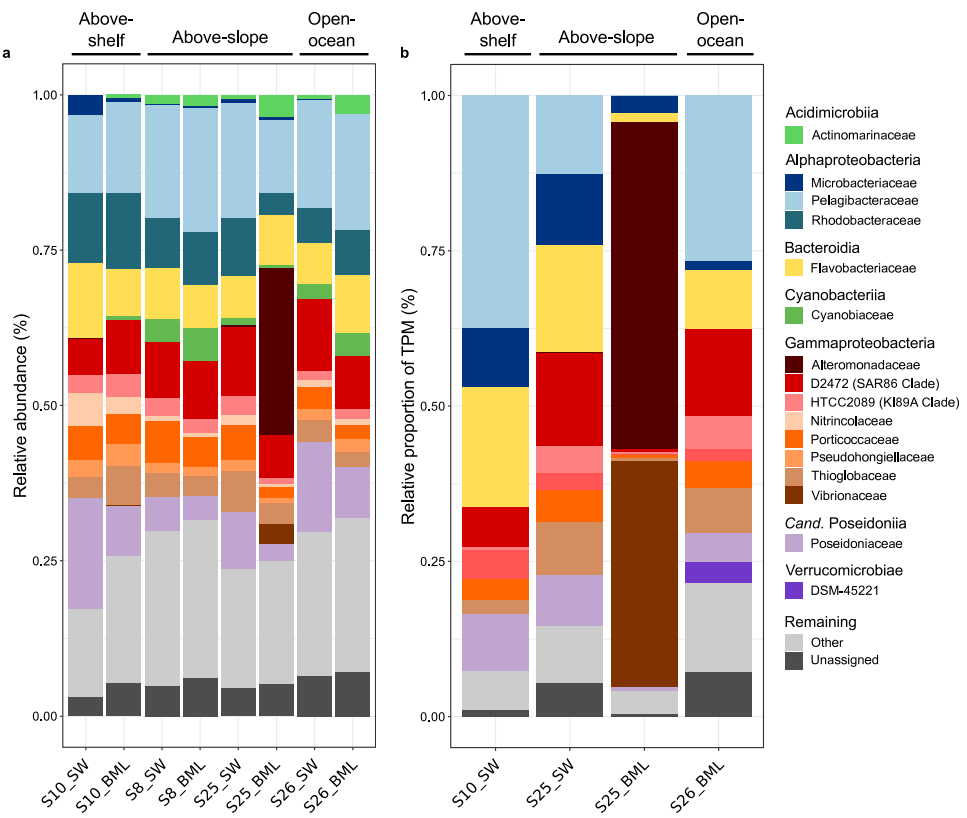
Supplementary Figure S4. Clustering analysis based on dissimilarity in sequence composition of metagenomes from the Arctic and high North Atlantic. Metagenomic sequence composition was compared using MASH. Metagenomes from this study, FRAM_WSC20, were compared to those previously sampled from the Fram Strait region, FRAM_RAS_EGC and FRAM18, along with those collected from the Arctic Ocean and high North Atlantic during the Tara Oceans expedition, Tara_.



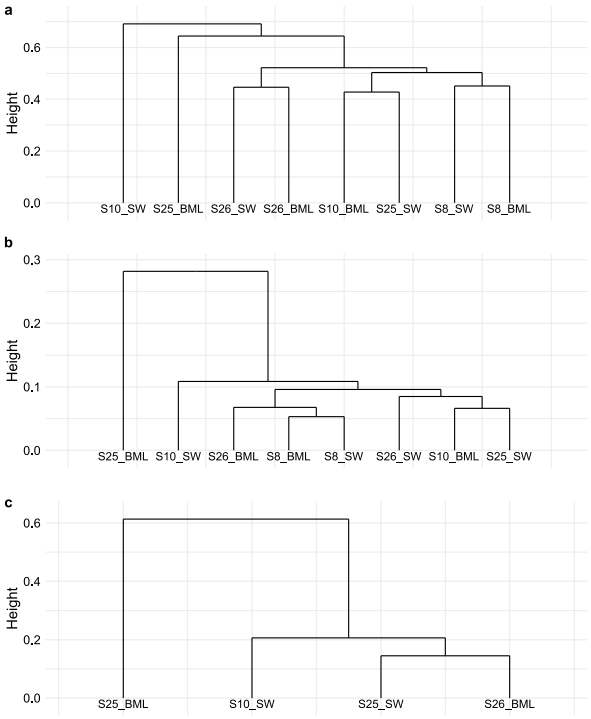
Supplementary Figure S5. Kingdom level taxonomic composition of metagenomes and statistics on the number of micorbila genomes and species sampled. a) Kingdom level composition of metagenomes, determined using Tiara. b) The number of microbial genomes and species captured was determined based on sequencing depth of single-copy ribosomal proteins.



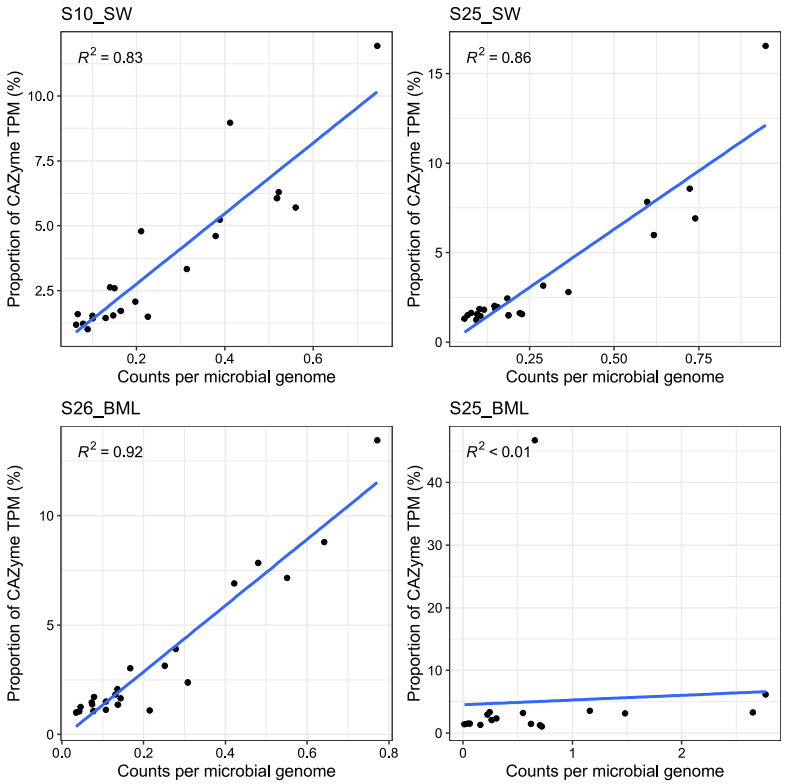
Supplementary Figure S6. Composition of eukaryotes in metagenome samples. Composition shown is based on 18S rRNA genes that were extracted from raw HiFi reads using Barrnap and analysed using DADA2.



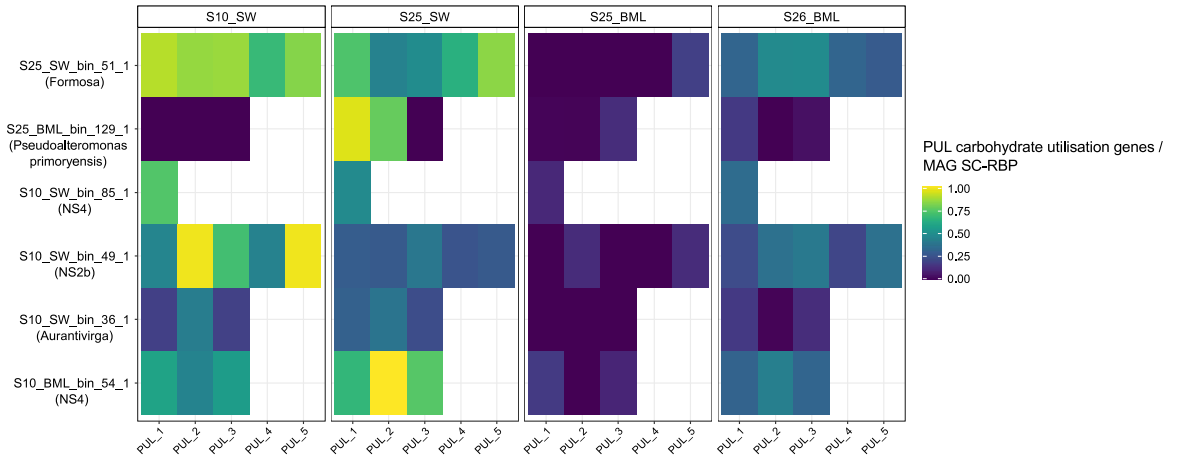
Supplementary Figure S7. Family level composition of microbial communities captured by the metagenomes and metatranscriptomes. Community composition was assessed using the large subunit ribosomal protein L6 gene, with the taxonomic assignment being derived from raw HiFi read classification against the GTDB database.



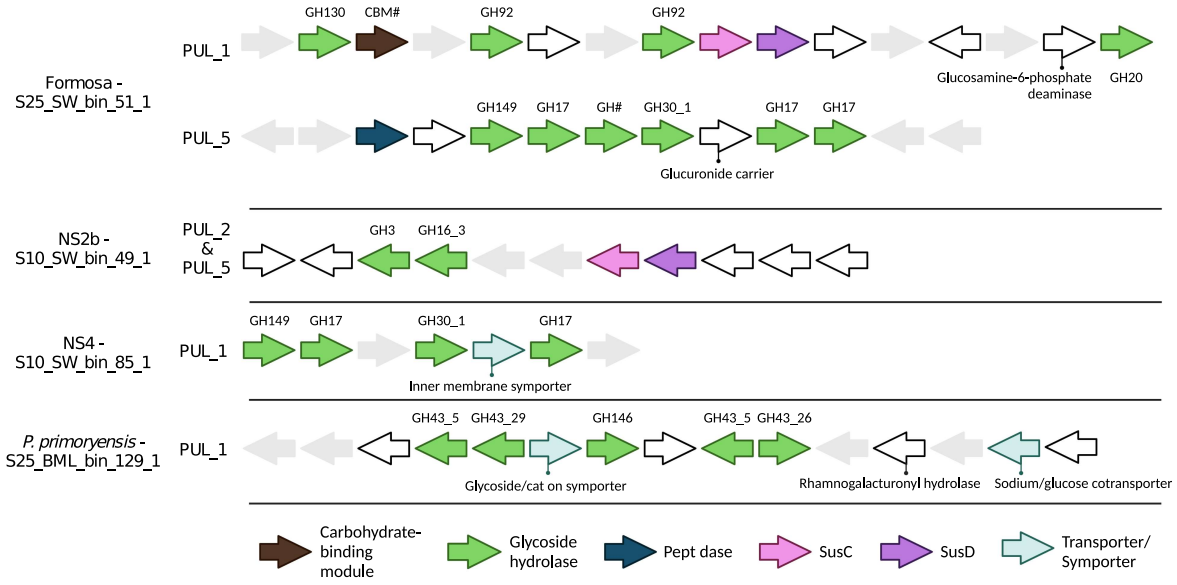
Supplementary Figure S8. Clustering analysis on taxonomic and functional composition of metagenomes and metatranscriptomes. a) Dissimilarity of microbila taxonomic composition of metagenomes, b) dissimilarity of microbial functional composition of metagenomes and c) dissimilarity of microbial functional composition of metatranscriptomes.



Supplementary Figure S9. Linear regression analysis of carbohydrate-active enzyme gene transcription against gene abundance. CAZyme gene abundance was normalised by the number of microbial genomes captured in the metatranscriptome (counts per microbial genome).



Supplementary Figure S10. Transcription level of carbohydrate utilisation genes located within polysaccharide utilisation loci (PULs). The selected MAGs represent those that contribute >5% to total MAG carbohydrate-active enzyme (CAZyme) gene transcription. The transcription levels shown were determined by taking the average TPM values of genes involved in carbohydrate utilisation within each PUL, including CAZyme genes, sulfatases and transporters. This average value was then subsequently divided by the average TPM value of single-copy ribosomal protein genes within each MAG, to a provide a comparison to the general genome transcription level.



Supplementary Figure S11. Visualisations of polysaccharide utilisation loci (PULs). The visualised PULs are those that were transcribed at levels comparable to, or higher than the MAG single-copy ribosomal protein genes. PUL diagrams were created on Biorender.com.

Chapter VI

Discussion

Discussion

6.1 The Fram Strait and Arctic Ocean microbiome

As a consequence of climate change and human-induced perturbations to the Earth's biosphere, biological communities and natural ecosystems are under tremendous pressure. In the Arctic, and surrounding geographical regions, a fundamental shift in the state of ecosystems is currently underway, driven by alterations to the hydrological cycle, expanding influence of inflowing Atlantic water, decreasing sea ice extent and high rates of atmospheric warming. Assessing how biological communities react to such changes is essential for understanding future ecosystem functioning in the Arctic. As the main entry point for Atlantic water and exit point for Arctic sea ice and freshwater, the Fram Strait is an invaluable region for studying such changes. Despite the chapters of this thesis each having different focal research questions, all the samples collected and generated for the work were derived from the Fram Strait over a period of 4 years (Figure 1). Agglomerating the observations across the datasets in conjunction with previous findings, we can provide detailed insights into the ecology of microbial communities inhabiting the Fram Strait. However, to really understand microbial ecology, these insights need to be placed into the context of oceanographic conditions and ecological observations of other trophic groups, such as primary producers. With this, we can provide a more rounded description on the ecosystem state and functioning and make predictions on how this may change under future conditions.

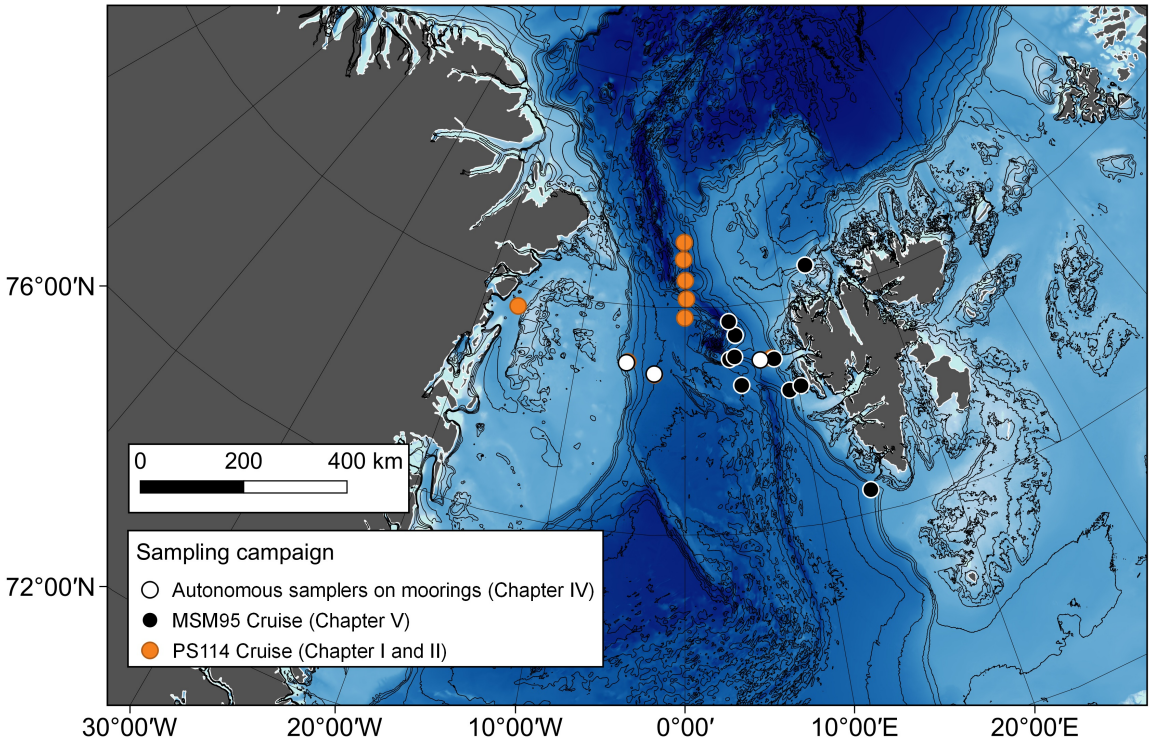


Figure 1. Locations in the Fram Strait that were sampled for the research in the thesis.

6.1.1 The current Fram Strait microbiome and its associated components

The complex and dynamic hydrography of the Fram Strait culminates into distinct physicochemical and biological regimes across longitudinal scales. The West Spitsbergen Current (WSC), which transports Atlantic water northward, maintains relatively stable year-round temperatures of above 4°C that prevents sea ice formation and thus preserves a temperate-like ecosystem at high latitudes. In contrast, the East Greenland Current (EGC), which transports polar water into the North Atlantic in the western Fram Strait, exhibits nearly year-round ice cover and temperatures of <0°C. As a consequence of these features and their associated effects on water column stratification, the EGC is considered more nutrient limited than the WSC [25] and subsequently exhibits lower primary productivity, <50 g C m^2 yr^{-1}. The high primary productivity of the WSC, 100 – 130 g C m^2 yr^{-1}, is associated with spring and autumnal phytoplankton blooms that are predominantly composed of diatoms [26], resembling temperate marine ecosystems. A similar distinction between ice-free, Atlantic waters and ice-covered polar waters in the Barents Sea (adjacent to the Fram Strait) has also been observed, with higher productivity and elevated contributions by diatoms under Atlantic conditions [27]. Phytoplankton phenology also varies across the Fram Strait, with blooms in the WSC typically occurring earlier (May – June) than in the EGC (July – August) [28]. Despite these broad trends, other studies from the Fram Strait have shown that the variations in biophysical factors drives high spatial variability in phytoplankton production, with maximum values being observed at the marginal ice zone [28, 29]. Thus, although conclusions can be derived from distinct regions and hydrographic conditions, the influence of local processes cannot be underestimated.

Congruent with differences in phytoplankton productivity and water mass characteristics, distinct spatial and temporal dynamics in organic matter concentration and composition have also been observed in the Fram Strait. Dissolved organic carbon (DOC) concentrations are typically elevated in the Arctic compared to other oceans due to a high influx of terrigenous and riverine DOC and slower degradation rates [30, 31]. These characteristics are preserved in the EGC, which exhibits higher DOC concentrations and elevated proportions of chromophoric dissolved organic matter (CDOM) than the WSC [32, 33]. Combining CDOM concentrations with modelled current velocities, Granskog et al. [34] estimated that up to 50% of the total terrigenous/riverine CDOM input to the Arctic is transported southward in the EGC. Comparable estimations have also been made from the analysis of ultra-filtered DOM [35]. In contrast, the concentration and composition of DOC in the WSC are intrinsically linked to phytoplankton production and exhibit strong temporal variations [36]. A 3.7% decrease in DOC was reported in the WSC between summer and

autumn in 2018, concurrent with a twofold reduction in dissolved carbohydrates and amino acids. Over the same time period, compositional changes in DOC also occurred, with increased bacterial-derived signals, such as amino sugars, and decreased phytoplankton-derived signals, such as neutral and acidic sugars, between summer and autumn [36]. These are patterns typical of temperate marine systems that experience rapid and pronounced impulses of organic matter from phytoplankton blooms in spring that is subsequently consumed by heterotrophic microbes.

In contrast to the DOC dynamics, the WSC exhibits a higher concentration of particulate organic carbon (POC) compared to the EGC, with enriched phytoplankton absorption properties [37]. A pan-Arctic investigation of POC concentrations spanning 25 years of sampling, reported the WSC to consistently contain the largest POC load [38]. Such patterns are not surprising, as POC is directly linked to phytoplankton production that, as already described above, is higher in this region. In the WSC, POC concentrations during summer are threefold higher than in autumn, with a corresponding decrease observed in cryptophyte and picoeukaryotic cell abundances over the same time period [39]. A congruent three-fold decrease in the proportion of POC derived from phytoplankton production was also observed between summer and autumn [39]. The autumn samples in that study were derived from September, the same period in the annual cycle that was sampled for Chapter V of this thesis. There, we evidenced high spatial heterogeneity in the carbohydrate fraction of particulate organic matter. Remnants of earlier productivity were consistently observed at all stations, through the presence of complex polysaccharides, such as sulphated fucans, that are known to accumulate in POM [40]. However, stations located above the continental slope also contained signatures of more recent production, with the presence of labile, simple structures, such as β-glucans. The spatial heterogeneity observed in the carbohydrate pool mirror those reported in phytoplankton productivity. Such spatial patterns are likely driven by the complex hydrographic dynamics in this region and again, highlight the importance of local processes. Furthermore, the discrepancies to von Jackowski et al. [39], indicate high inter-annual variability.

The fundamental distinctions in primary productivity and nutrient regimes along with the dynamic hydrographic and physical conditions could be expected to act as strong selective pressures on microbial communities, culminating in unique assemblages in eastern and western regions of the Fram Strait. Surprisingly though, work in this thesis and previous studies have reported that a large proportion of the microbial community, in terms of relative abundance, persists, irrespective of conditions. In Chapter IV, we identified 232 amplicon sequence variants that consistently represented the largest proportion (48 – 88%) of the microbial community regardless of environmental conditions, which we suggested was the resident microbiome of the Fram Strait (Figure 2). This finding is in agreement with Fadeev et

al. [41], who observed a minor fraction of the microbial diversity (13%) consistently dominated communities (75% on average) during summertime across a longitudinal transect in the Fram Strait. To this author, this observation is quite remarkable. It suggests that the resident microbiome may be locally adapted and highly flexible, being composed of populations suited for all extremes. Changes in light (polar day to polar night), productivity, nutrient regimes and physicochemical conditions are thus responded to by structural shifts in the microbiome. In Chapter IV, the observed resident microbiome encompassed a large phylogenetic diversity, with 79 distinct genera, that would provide the spectrum of phenotypic and metabolic traits to support such a concept. The populations that we identified as reaching the largest relative abundances within the resident microbiome were affiliated with *Candidatus* Nitrosopumilus, *Polaribacter,* Arctic97B-4 and the SAR11, SUP05 and SAR86 clades (Figure 2), which are congruent with those reported by Fadeev et al. [41]. The presence of a resident microbial community that comprises a minor fraction of diversity but a major fraction of relative abundance is a feature frequently reported in marine pelagic ecosystems [42], but previous studies were typically focused on temperate sites that do not experience such dramatic shifts in environmental conditions.

Despite the persistent resident microbiome in the Fram Strait, high influxes of Atlantic water can result in shifting community dominance to chemoheterotrophic populations that taxonomically resemble, at the genus-level, those in coastal temperate ecosystems. These populations were taxonomically affiliated with known heterotrophic groups, such as *Luteolibacter* and *Flavobacterium*, and typically peaked in response to phytoplankton bloom events (Chapter IV). Rapid growth of heterotrophic clades in response to phytoplankton blooms or nutrient influx events are well evidenced in temperate coastal ecosystems [43, 44], wherein they can shift from <1% relative abundance to dominance of the microbial community within a week. Such growth dynamics are evidence of microbial responses to the opening of ecological niches – this will be addressed more thoroughly later in this discussion.

The major proportion of phylogenetic diversity in microbial communities of the Fram Strait was observed in the ‘rare’ fraction, composed of populations that never surpassed 0.1% relative abundance. In the amplicon-based analysis in chapter IV, this fraction constituted 70% of the diversity. The persistent nature of rarer microbial populations is also a feature previously observed in other areas of the Arctic Ocean [45] and in coastal temperate regions [46]. The ecological role of the rare community fraction is a topic that is currently under debate. However, it is likely that these populations include those existing in a state of dormancy, acting as seed populations ready to respond to changes in conditions, and those that are

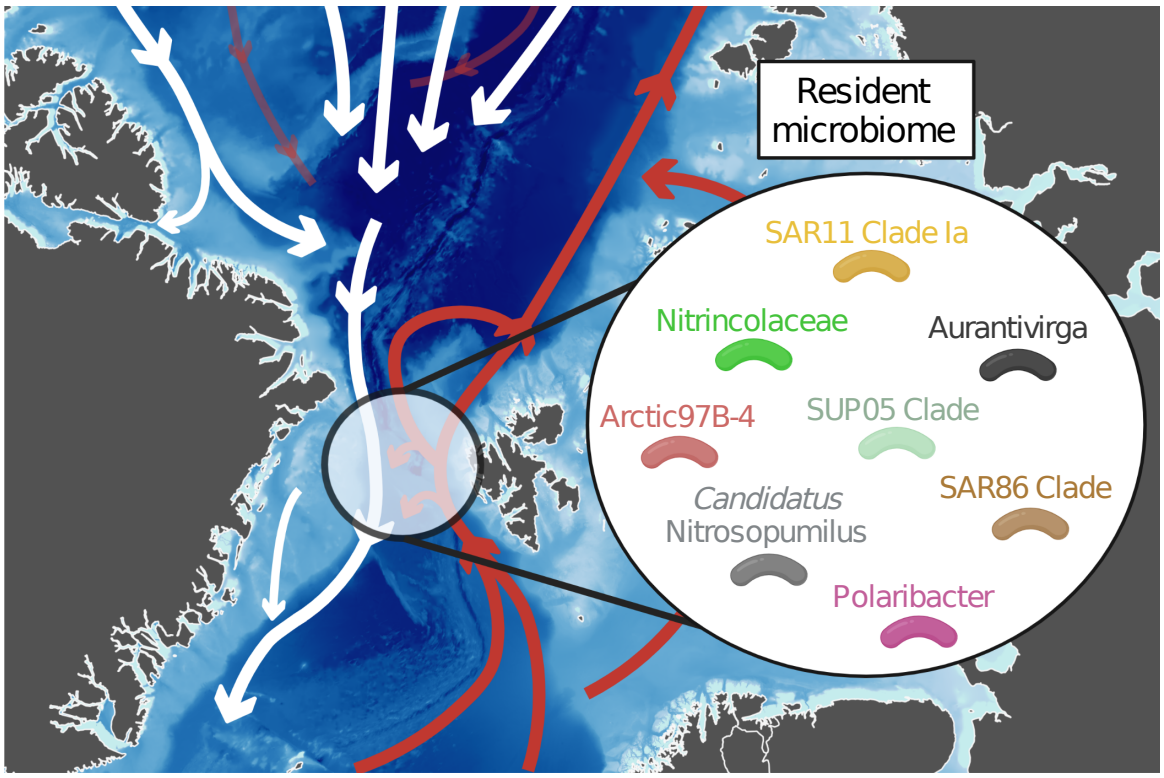


Figure 2. Map of Fram Strait with a conceptual representation of the resident microbiome. The resident microbiome identified in Chapter IV was comprised of 232 amplicon sequence variants that were consistently present over a four year time-series, regardless of conditions. The taxonomic groups illustrated in this figure are those which the 15 most abundant amplicon sequence variants were assigned to.

metabolically active and responsive but filling ecological niches that do not permit expansion to large population sizes.

In addition to assessing temporal dynamics in microbial communities, we also observed distinct shifts in composition and structure over spatial scales within the eastern Fram Strait in Chapter V. Despite the samples for that project being collected at similar times (within two weeks of each other) and in water masses with similar physicochemical properties (predominantly north Atlantic water conditions), microbial assemblages shifted at spatial scales of <100 km. This patchiness highlights the responsiveness of populations within microbial communities and the importance of local processes in shaping community structure. Such observations have been observed previously at even smaller spatial scales (<10 km) [47]. Whether or not the observed spatial dynamics were driven by members of the resident Fram Strait microbiome or those being introduced by inflowing Atlantic water is yet undetermined. However, the datasets from Chapter IV and V are currently being agglomerated in order to make such assessments.

In conclusion, the Fram Strait is a highly dynamic oceanographic region with extreme variations in conditions occurring over temporal and spatial scales that influence and shape biological communities. These conditions range from seasonal transformations in ice-cover and daylight down to biophysical processes that occur on a local scale within water masses.

With the current trends in climate warming however, dramatic perturbations are taking place in the Fram Strait and wider Arctic Ocean. Understanding the impact these have on biological communities is of paramount interest, as it can help to predict future shifts in the state and functioning of ecosystems.

6.1.2 Future shifts in the Fram Strait and Arctic Ocean microbiome and their associated components

Among the major perturbations occurring in the Arctic Ocean as a consequence of climate warming are a reduction in sea ice extent and thickness, increase in terrestrial and riverine input, acceleration of coastal erosion rates and an expansion of Atlantic water influence. Such processes impact all components of ecosystems from the hydrography to physiochemical state and organic matter content and thus, will have a profound effect on biological communities. A number of key observations have already documented shifting patterns in primary producers and higher trophic levels in response to changing conditions. However, the research in this thesis was amongst the first to assess microbial dynamics over spatial and temporal gradients in Arctic – Atlantic conditions and make predictions on future shifts in the Arctic Ocean microbiome. In the following section, the findings from this thesis will be combined with previous work to provide an overview on the current trends in the changing Arctic Ocean and Fram Strait.

Net primary production (NPP) across the Arctic Ocean has increased significantly in recent decades. Although the reported magnitude of change in NPP differs between studies, the most recent reports indicate an overall increase of 57% since 1998 [48]. However, the patterns were not spatially uniform and high regional variability was observed. The largest increases were detected above the inflow shelves, Barents and Chuckhi, and eastern interior shelves, such as Laptev, with the inflow shelves contributing 70% to the total Arctic Ocean NPP increase. Prior to 2009, increased NPP over interior shelves was attributed to reduced sea ice extent that facilitates a longer growing season and expanded habitat space for pelagic phytoplankton [49]. However, since 2009, the observed increase has been positively correlated with phytoplankton biomass whilst only minor changes to sea ice extent and open water habitat space have occurred [48]. In the Barents region and Eurasian Arctic, elevated phytoplankton biomass is likely a result of an expanding Atlantic influence [50] that results in higher nutrient availability. Inflowing Atlantic water is more nutrient rich than polar water, particularly with respect to nitrate that is considered a limiting factor in the Arctic Ocean. In addition, the increased heat content of Atlantic water accelerates sea ice decline, which reduces stratification and facilitates deeper vertical mixing and the resupply of nutrients from depth [50]. In the eastern Fram Strait, changes in the composition of phytoplankton

communities has also been observed, with a shift in dominance from diatoms to *Phaeocystis* [51]. The northward expansion of temperate species' habitat space is also underway, which will add increased biological pressures to Arctic communities on top of the hydrographic and physicochemical changes. The above-described shifts occurring at the base of the food web will have cascading effects, altering nutrient regimes and organic matter composition and availability. The availability of organic matter in the upper ocean is also directly influenced by sea ice dynamics. Sea ice derived meltwater stratification supports longer lasting phytoplankton blooms that retain organic material in the surface water and ultimately slow the biological carbon pump [52]. In contrast, areas that are unaffected by sea ice, exhibit pronounced, short-lived phytoplankton blooms that result in large pulses of organic carbon export to the deep. As microbes are reliant on organic matter for growth, reducing sea ice extent and shifts in phytoplankton communities will likely be major factors that reshape microbial communities in the future.

Indifference to NPP and phytoplankton communities, there has been no clear evidence on how expanding Atlantic influence is impacting microbial communities. However, in Chapter IV of this thesis, we leveraged a time-series dataset incorporating two distinct sampling locations to characterise microbial communities under different hydrographic regimes and ecosystem states that can provide a foundation for predicting future shifts. We observed that with high Atlantic water influx, microbial communities were dominated heterotrophic populations that taxonomically resemble those of temperate ecosystems, such as the NS5 Marine group and *Amylibacter* (Figure 3). This pattern reflects previous findings where distinct water masses have been shown to harbour unique microbial taxonomic signatures [53, 54] and indicates that temperate microbial populations will likely become more widespread with expanding influence of Atlantic water in the Arctic ocean. Although we were unable to say whether such a process would result in the complete loss of specific populations or species, 155 amplicon sequence variants exhibited strong negative associations with Atlantic water and positive associations with polar water. For these populations, expanding influence of Atlantic water will likely diminish their ecological niche space, thus reducing the area over which they can be competitive. Such a process is likely also not restricted to the Eurasian Arctic, where Atlantic water exerts the most influence. The other major entry point for temperate oceanic water into the Arctic Ocean is the Bering Strait, where the inflow of Pacific water takes place. Between 2001 to 2015, an almost twofold increase in the volume of Pacific water flowing through the Bering Strait was observed, whilst heat and freshwater fluxes concurrently increased [55]. As of yet, only limited studies have assessed microbial communities in this region and thus it is inconclusive whether a similar northward expansion of temperate-like populations will occur. However, if the Pacific inflow is increasing then it is

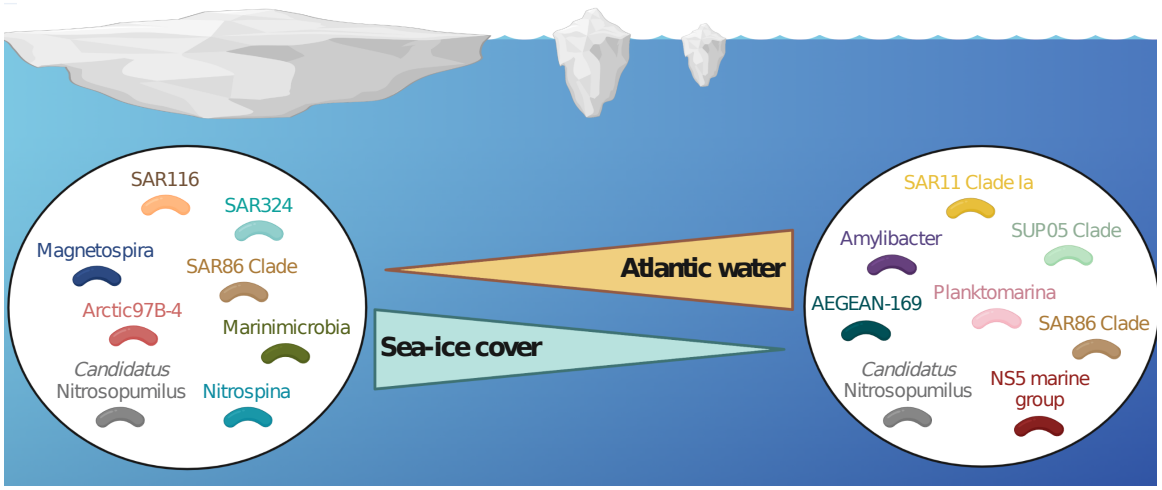


Figure 3. Conceptual diagram illustrating the predicted changes in the resident microbiome of the Fram Strait and Arctic Ocean under different conditions. The selected taxonomic groups are based on the affiliation of the most dominant amplicon sequence variants under the two contrasting conditions observed in Chapter IV. Where the same clade occurs in both scenarios, the shift occurs at a higher phylogenetic resolution.

conceivable that it will shift the hydrographic properties in favour of Pacific-derived microbial populations, similar to what we observed with Atlantic inflow.

Another major perturbation that will drive fundamental shifts in Arctic marine microbial communities is reducing sea ice extent. In Chapter IV, we were able to show that sea ice cover had the largest impact on community functionality out of the tested variables, with low ice cover resulting in enriched metabolic signatures for phytoplankton-derived organic compound degradation compared to inorganic substrate and bacterial-derived compound degradation under high ice cover. It is important to keep in mind that we were only able to assess whole community functionality for a limited portion of the dataset in chapter IV, and thus much more extensive analysis would be required to further corroborate our findings. However, considering how reduced sea ice extent has been linked to increased NPP and phytoplankton biomass, our observations are likely good indicators. In addition to the already described impacts of sea ice on biological communities, sea ice also plays a major role in light attenuation (UV radiation), sea-air gas exchange and the overall climate feedback loop in the Arctic, each of which exert different pressures on microbial communities. UV radiation is known to chemically alter dissolved organic matter (DOM) through the process of photooxidation, which can change its composition and release dissolved inorganic carbon. A study by Bélanger et al. [56] evidenced that reduced sea ice cover can increase the proportion of photochemically altered dissolved organic carbon, particularly that derived from riverine input. The reduction of air-sea gas exchange under high ice cover can also result in trapping of gases in the water column, which has been evidenced in the case of methane. Methane seepage from sub-seafloor gas hydrates has been evidenced at a number of locations across Arctic shelf regions [57, 58]. An Arctic-wide study on methane fluxes from sediment to the atmosphere, revealed that methane

reached super saturation under high ice cover but coincided with high microbial methane oxidation rates [57]. In contrast, low ice cover resulted in reduced water column methane concentrations and lower methane oxidation rates, likely resulting from shorter residence time in the water column and increased exchange with the atmosphere. Notwithstanding the consequences of these findings from a climate perspective, the change in saturation and reduced oxidation rates in the water column highlights another direct impact that a reduction in sea ice extent will have on microbial communities. The so far described consequences have largely been focused on pelagic microbial communities but it is important to note that sea ice itself harbours a rich phylogenetic and functional diversity of microbial populations. At this point I would draw the reader's attention to several studies that have assessed the microbial communities inhabiting sea ice [59–62], as it is beyond the scope of this discussion. It should thus be clear that sea ice has numerous direct and indirect impacts on microbial communities, and its continued loss in the future will progressively reshape the composition and structure of microbial communities across the Arctic Ocean.

From our observations in Chapter IV, we predicted that the combination of reduced sea ice extent and increased Atlantic water influence will result in compositional and functional changes in microbial communities. Notably, microbial communities that resemble those of temperate ecosystems and exhibit pronounced seasonal dynamics will become more widespread whilst those adapted to Arctic conditions, such as high-ice cover, will likely be restricted to the central Arctic Ocean and EGC. A conceptual diagram illustrating some of the compositional shifts that may occur is provided (Figure 3). This illustration however, does not incorporate those groups that respond to seasonal phenomena such as phytoplankton blooms, which in chapter IV consisted of *Luteolibacter*, *Aurantivirga, Ulvibacter* and *Lentimonas* under high Atlantic water and low sea ice conditions.

The input of organic matter and freshwater into the Arctic Ocean is also drastically changing as a result of permafrost thawing, glacial melt and coastal erosion. Permafrost-rich Arctic coastlines constitute 34% of the world's coastlines [63]. Typically in the Arctic, these coastlines are protected by landfast- and sea ice for extended periods in the year that reduce their exposure to wave action and thus erosion. With reducing sea ice extent, coastal erosion rates are increasing. In parallel to this, the permafrost along the coastlines is thawing due to increasing air temperature, which further expedites coastal erosion. In regions of Alaska, coastal erosion rates of up to 25 m yr^{-1} [64] have been reported. These processes result in tremendous quantities of organic matter being flushed into the Arctic Ocean, with an estimated annual flux of particulate organic carbon of 14 Tg [65, 66]. The exact fate of this organic matter is still under investigation, but initial evidence indicates a high rate of deposition into nearshore sediments [67]. However, we hypothesise that this increasing terrigenous organic matter influx to the Arctic Ocean is bound to have a direct influence on microbial communities. A recent

study from the Canadian basin recovered genomes of *Chloroflexi* that are enriched in aromatic compound degradation genes and functionally distinct from other marine *Chloroflexi* [68]. The authors of that study further indicated that the genes were laterally transferred from terrestrial microbes. This suggests that some microbial clades may be able to adapt to and make use of the increased terrigenous organic matter load. A more thorough investigation of this is currently underway and formulates my next major research project, where we will combine organic matter characterisation with metagenomics and metatranscriptomics along transects of the Beaufort Sea Shelf to determine how terrigenous organic matter input influences microbial communities.

6.1.3 Outlook and directions for future research on the Fram Strait and Arctic Ocean microbiome

The research conducted throughout this thesis provided valuable and novel insights into the Fram Strait and Arctic Ocean microbiome, particularly with respect to dynamics over temporal scales and in association with ice cover and water masses. From our findings, we were able to make predictions on how microbial communities may change under future conditions in the Arctic Ocean. In general, we propose a shift to a more temperate-like ecosystem state, where biannual phytoplankton blooms are observed and ice-influenced processes are largely restricted to the central basin and EGC in the Fram Strait. Concurrently, microbial communities will become dominated by heterotrophic members that taxonomically and functionally resemble those of temperate ecosystems and exhibit pronounced seasonal dynamics. As a consequence of this, microbes adapted to Arctic conditions will likely only be able to remain competitive in the central Arctic Ocean and in the EGC. However, there still exist many gaps in our knowledge and further research is necessary to validate our predictions. Most notably, is the lack of data on microbial communities inhabiting the central Arctic Ocean, which has persisted due to the logistical challenges of sampling in sea ice. As the EGC is supplied by the Transpolar Drift, we believe that our ecological observations will likely be representative for central Arctic Ocean microbial communities but this needs to be confirmed. The recent completion of the largest ever Arctic sampling campaign, MOSAIC, will provide the data to test this and deliver the first temporal dataset from the central Arctic Ocean. With the data being generated from that expedition, we will be able to make more thorough ecological assessments and more accurately understand how changing conditions are impacting microbial communities in the Arctic Ocean.

6.2 Long-read metagenomics as a tool for investigating microbial ecology

Over the past decade, the application of next generation sequencing platforms in microbial ecology research has resulted in the recovery of hundreds of thousands of genomes representing environmental populations. This has not only tremendously expanded the tree of life but also allowed for functional descriptions of taxonomic groups for which our ecological information was limited or non-existent. Despite this, we were still only able to characterise a fraction of the community, with a large proportion of reads not being accounted for in assemblies or the recovered metagenome-assembled genomes (MAGs). Third-generation sequence (TGS) platforms present opportunities to overcome these limitations and gain a more complete overview on a sampled microbial community. In particular, HiFi reads generated from the PacBio Sequel II and IIe platforms that boast a per base accuracy of >99% and average length >5 kbp. HiFi read metagenomics formulated the primary investigative tool employed throughout the work of this thesis. In chapter II, we showed an improved quality and increased number of MAGs could be recovered using HiFi over short-read metagenomes. However, it is the capacity to analyse community taxonomy and function directly from the raw HiFi reads that really showcases the distinct advantage of TGS over NGS technologies. This approach was taken in chapter IV and V of this thesis. However, due to the novelty of using HiFi read metagenomes to investigate environmental microbial communities, a tremendous amount of work over the course of this thesis was dedicated to testing, modifying and optimising pipelines and different approaches to most efficiently and effectively gain ecological information from such datasets. As a result, in the following section I aim to summarise some of the key developments made whilst proposing novel metrics and ways to assess microbial communities using long-read metagenomes.

6.2.1 Assessing phylogenetic composition of microbial communities in long-read metagenomes

Obtaining taxonomic information on the genetic content of microbial communities is essential for answering our core ecological question, “What microbes are there?”. Traditionally, this aspect was addressed by the sequencing of phylogenetic marker genes, such as the 16S rRNA gene (hereon referred to as 16S amplicons). Still, to this day, amplicon sequencing is widely used and far from being an obsolete approach, particularly with the tremendous reduction in sequencing costs and the relative ease of data analysis. With long-read metagenomics however, there are several different avenues that can be used for assessing diversity and composition in microbial communities. The following sections will highlight different approaches used throughout this thesis.

6.2.1.1 16S rRNA gene-based approach

The 16S rRNA gene has long been the standard unit of phylogeny for microbes due to its universality and variable degree of conservatism across its length, amongst other reasons [19, 69]. Its wide-scale application in marine microbial ecology has resulted in large, highly curated databases being developed concurrent with tools and methods for analysis of the data [70–72]. In order to make use of such valuable resources and allow for comparative analysis to previous studies, a 16S rRNA gene based approach can also be employed to study diversity and composition in long-read metagenomics, with some additional benefits to amplicons.

With the complete rRNA gene operon length ranging from 4 – 6 kbp, the rate of recovery of complete 16S rRNA genes within single long-read sequences is high. In chapter II, we extracted >17,000 full-length 16S rRNA genes from three HiFi read metagenomes. In contrast to the partial genes captured with amplicon sequencing, having access to the full-length gene has added advantages. Notably, an increase in the resolution of phylogenetic classification and discrimination at lower taxonomic levels. Furthermore, it provides more opportunity to design highly specific probes that can be used in fluorescence *in situ* hybridization for visualization of populations in the environment.

Despite its long-standing and wide-scale use, there exists a number of limitations and challenges with using the 16S rRNA gene for ecological studies on microbial communities. Namely, 1) the phylogenetic resolution below the genus-level is limited, 2) most microbial genomes harbour several rRNA gene operons [73], 3) intra-genomic variability across operons and inter-genomic variability between species are not always distinguishable on a sequence similarity level, and 4) explanatory power between lineages is highly variable [74]. A number of recent studies, in an effort to circumvent these limitations, have compared and proposed alternative universal marker genes that provide better discriminatory power than the 16S rRNA gene at lower taxonomic levels [75, 76]. Promising candidates that were used in chapter V of this thesis, are single-copy ribosomal protein genes.

6.2.1.2 Ribosomal protein gene-based approach

Ribosomal protein genes (RBP) have been shown to harbour considerable power for phylogenetic analysis. Through the use of 16 single-copy RBP genes, Hug et al. [77] presented a significantly expanded version of the tree of life and evidenced underrepresented lineages that contained major evolutionary radiation. Although the high conservatism within RBP genes allows for deep phylogenetic relationships to be resolved, a recent study by Olm et al. [78] also concluded that they offer a high species discrimination power. Comparing numerous metrics to verify and delineate bacterial species within metagenomic datasets, Olm et al. identified several RBP genes that could serve as better candidates for species-delineation than other typically used marker genes based on recoverability and species

delineation accuracy. Using >5000 bacterial genomes from environmental metagenomes, Olm et al. further reported species-delineating average nucleotide identity thresholds for several RBP genes. As a result, RBP genes can be used to reconstruct phylogenetic trees that include distant and recent evolutionary divergences and allow for the identification of species, which provides the fundamental unit from which diversity can be measured. Furthermore, the single-copy nature of many RBP genes allows for more accurate estimations of species relative abundances and thus opportunities to better assess community structure and diversity.

In chapter V of this thesis, we evidenced how RBP genes can be used to characterise the phylogenetic composition of microbial communities and gain insights into diversity and complexity. We leveraged the four RBP genes proposed by Olm et al., whose species-delineating ANI thresholds exhibited the highest discrimination accuracy, to determine the number of species in each metagenome and subsequently obtain abundances through read mapping. By normalizing the number of species by sequencing effort (per million reads), we can thus gain an insight into the diversity and complexity of a sampled community. Employing this approach in chapter V, we evidenced a twofold difference in the normalised number of species between samples. I believe that such an approach could be adopted in future long-read metagenomic studies and provide a metric for comparing community diversity and complexity between datasets. Another valuable improvement that can be made using RBP genes is a more accurate assessment of community structure. In amplicon-based studies, the structure of communities is typically presented through relative compositions. However, such an approach is prone to error and likely inaccurate owing to the fact that >80% of microbial genomes harbour multiple 16S rRNA gene copies [73], copy number varies across phylogenetic lineages and the primers used to amplify the 16S rRNA gene are known to be biased. With RBP genes however, only ~1.5% of bacterial genomes are identified as harbouring more than one copy [78]. In chapter IV and V, we calculated the total number of genomes captured in each metagenome, based on the average sequencing depth of 16 single-copy RBP genes. This provides the essential ‘total’ value from which relative proportions of species can be determined. By determining the sequencing depth of single-copy RBP genes for each species in relation to this total, it provides a more accurate representation of community composition and structure. Furthermore, in chapter V we proposed a novel metric, number of genomes per species per million reads (GPSM), which could be employed in future long-read metagenomic analyses to allow for comparisons of species richness and evenness across samples and datasets.

As with all techniques we use in microbial ecology, assessing community composition with single-copy RBP genes is not without limitations. The major difficulty is with assigning taxonomy. Unlike the 16S rRNA gene, there are no large, comprehensive databases for single-copy RBP gene phylogeny. As such, in chapter IV and V, a novel pipeline was

employed. This pipeline is based on the taxonomic classification of entire long reads, from which the single-copy RBP gene classification can be derived.

6.2.1.3 Whole read phylogenetic classification

Owing to the length and quality of HiFi reads, it is possible to perform taxonomic classification directly at the raw read level. In addition to providing information on community composition, this approach also allows for the valuable connection between phylogeny and function to be made. Performing such a process requires careful consideration however, particularly with respect to the databases used. A number of tools have been developed for the taxonomic classification of metagenomic contigs (assembled short-reads) that are applicable for long reads, such as CAT [79] and MMseqs2 taxonomy [80]. Although these tools vary in their methodology, they are typically reliant on gene prediction using Prodigal [81] and the NCBI-nr database as a reference. Prodigal was developed for short-read metagenomic sequences and lacks the ability to detect frameshift mutations that are commonly observed in long reads from TGS platforms. As such, frameshift aware gene prediction tools, such as fraggenescan [82], are necessary to prevent genes being spliced. With respect to databases, it is this authors opinion that the Genome Taxonomy Database (GTDB) [83] is a considerably more powerful resource for metagenomic read classification than the NCBI-nr database. The GTDB was established to standardise microbial taxonomy and is widely used as the reference database for phylogenetic classification of MAGs. The advantage of using GTDB is that it incorporates published MAGs from environmental metagenomes into its reference phylogenetic framework, assigning each a classification down to the species-level and designating new taxonomic groups where necessary. As such, classifying metagenomic reads against the GTDB will provide a higher resolution of classification for groups with uncultured representatives and allow for comparative analysis to other metagenomic studies. The most recent GTDB database release (r207.2) houses >317,000 genomes assigned to >62,000 species. With this in mind, in chapter IV and V of this thesis, we employed a custom long-read classification pipeline that was inspired by the framework of CAT but used the GTDB database as a reference.

In chapter V, we showcased that the custom GTDB-based pipeline can provide a wealth of taxonomic information. Employing the pipeline to the eight PacBio HiFi read metagenomes resulted in >99% of reads being assigned to a kingdom, 70% to a family and 43% to a genus. Despite the large drop in classification rate between the family and genus level, these values are comparable, if not higher than those typically observed in classifying operational taxonomic units (OTUs) of 16S rRNA genes [84]. Furthermore, considering that these classifications cover the entire genetic content of the community and not a single, extensively studied marker gene, further emphasises the value of the pipeline used.

Since the development and application of the pipeline used in chapter IV and V, additional new tools have been published that are specifically designed for long read classification. A recent comparison of such tools was made by Portik et al. [85], with MEGAN-LR [86] exhibiting the highest classification rate to precision ratio in the studied mock communities whilst also resulting in the least number of false positives. Unfortunately, the MEGAN-LR pipeline developed for that study was also reliant on the NCBI-nr database and the comparison on mock communities does not reflect the effectiveness of such tools on marine metagenomic samples.

6.2.2 Outlook and directions for future research in applying long read metagenomics in marine microbial ecology research

In comparison to their short read counterparts, long read metagenomes offer a number of distinct advantages for investigating microbial ecology. Most notably, is the capacity to assess phylogeny and function for the entire sequenced community, irrespective of the genomes that can be recovered with computational tools. This can facilitate investigations on the ecological roles of less abundant microbial populations and more accurate assessments on community structure and organisation. However, owing to their novelty, there is currently a lack of tools specifically designed for analysing long read datasets. In this thesis, a number of pipelines and methods were tested and new approaches were proposed for gaining insights into microbial ecology using long-read metagenomes, but further developments are required and validations through comparative analyses need to be made. For example, I proposed that employing a single-copy RBP gene-based approach may provide more accurate representations of community structure. This hypothesis can be tested by performing a detailed comparison on the relative proportion of taxonomic groups using cell counts (fluorescence *in situ* hybridization), single-copy RBP and 16S rRNA genes. With respect to the taxonomic classification of long read sequences, I believe that the pipeline employed in chapter IV and V could be further improved by incorporating all 317,000 genomes of the GTDB database as a reference as opposed to the 62,000 species representatives that were used. In addition, that pipeline needs to be compared to recently developed tools such as MEGAN-LR [86] and SprayNPray [87] to determine its effectiveness and accuracy.

Returning to the question raised in the introduction, is the higher cost of long read sequences suitably compensated for by the ecological insights that can be gained? Hopefully, the research conducted in this thesis along with the topics discussed above has provided a foundation and a point of reference for which that question can be answered in future projects. It is clear that the high cost is a major barrier preventing wide-scale adoption and thus in my opinion, long read sequencing is currently only suitable for certain projects. Naturally, as new

TGS platforms are developed, costs will decrease and I imagine that in five years, long read metagenomics will replace short read metagenomics as a primary investigative tool in marine microbial ecology.

6.3 Ecological niche concept

Since its conceptualisation more than 100 years ago, the niche concept has become a core tenet of ecology. Although its origins can be traced back to the seminal works by Grinnell [88], and later by Elton [89], the concept that is most often referred to today is derived from G. Evelyn Hutchinson, who made considerable developments on the earlier theories. Hutchinson proposed that a species' or population's niche is an *n*-dimensional volume defined within a hyperspace where the conditions and resources of an environment, also termed the species' requirements, are the axes [90, 91]. Interaction between species, e.g. competition, was also considered and the proposed model was defined as the "fundamental niche". Hutchinson further described that a fundamental niche could only be occupied by a single species within an environment, such that if two species occupy the same niche, only one will prevail. Although the work of Hutchinson and others at the time was based on macroorganisms, the niche concept was more recently adopted in the field of microbial ecology.

The plethora of research conducted on marine microbial communities in recent decades has resulted in two conclusions, a) microbial communities are highly dynamic and in a state of constant flux, and b) consistent patterns over spatial and temporal scales indicate that, to some degree, microbial dynamics are deterministic. The deterministic nature of microbial dynamics underpins their adaptation to certain conditions and thus supports the niche concept. Deciphering the ecological niches of marine microbes however, is a challenging task due to the myriad biotic and abiotic conditions that vary over large magnitudes of scale. Nonetheless, work conducted throughout this thesis along with previous studies have provided valuable insights into how certain conditions may influence th3e dynamics of microbes and thus contribute to defining ecological niche space.

6.3.1 Factors that contribute to shaping ecological niches

In this section I am to provide a summary on the niche-shaping conditions that were identified in this thesis and discuss their influence with respect to spatial and temporal scales. Following that, I will briefly consider other important factors that were not address in this thesis, whilst proposing a new modified niche concept that may be more applicable for marine microbes.

6.3.1.1 Temperature and salinity

Studies investigating microbial biogeography and community composition over large spatial scales have evidenced distinct patterns in microbial distribution across latitudinal and depth [92] gradients as well as with distance from a coast. These broad-level patterns reflect changes in distinct physicochemical conditions that occur across such gradients. In particular, it has been shown that temperature [93, 94] and salinity [95] strongly correlate with changes in microbial community composition and thus could influence microbial distribution. Despite these findings, it is this author's opinion that salinity likely has little direct impact in driving microbial distribution in oceanic environments, but instead plays an important role at boundary regions that experience large freshwater input. Studies assessing microbial communities along natural salinity gradients such as river estuaries and the Baltic Sea have highlighted a threshold of 5 – 8 psu that separates freshwater from marine compositions [95, 96]. Such findings have also been supported by cultivation-based efforts using marine isolates [97]. As all major oceans exhibit salinity values well above this threshold (typically >25 psu), the direct impact of salinity changes on microbes in the oceans is likely small. However, salinity can have an indirect impact on microbial distribution. Sharp changes in salinity and temperature result in density gradients that can act as physical barriers to dispersal for microbes; these exist most commonly in vertically stratified water bodies and horizontally between two distinct water masses, such as the polar- and Atlantic-derived water masses of the Fram Strait.

In contrast to salinity, temperature is expected to be a key determinant of microbial distribution and thus contribute to shaping ecological niche space. Cultivation-based investigations have long shown that microbes optimally grow within a defined range of temperatures as a result of genetic and phenotypic adaptions, e.g. mesophiles that grow

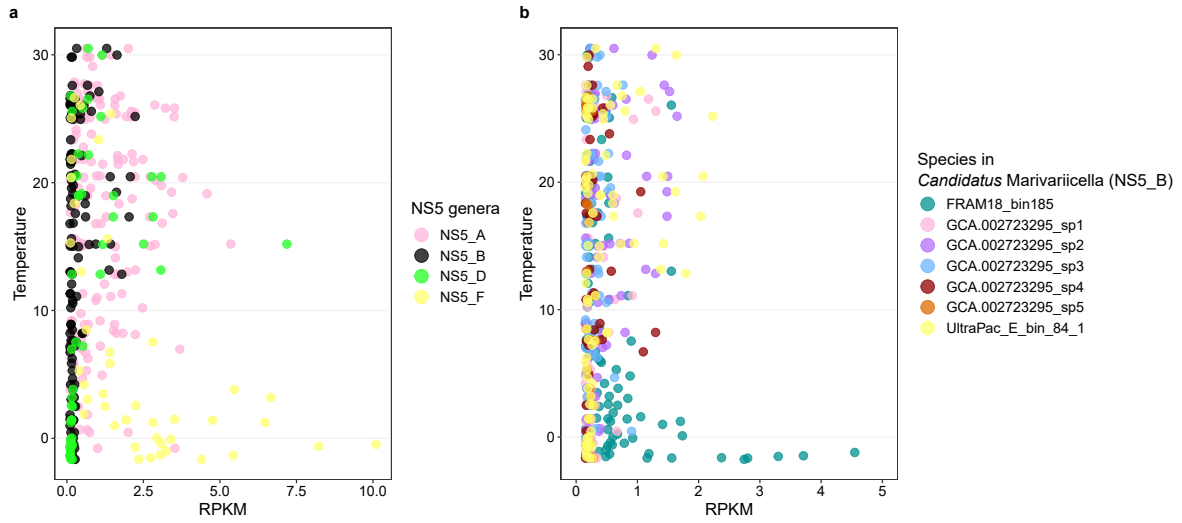


Figure 4. Distribution of NS5 genera and species in relation to temperature across the global Tara Oceans metagenomic dataset. Distribution is based on RPKM values calculated from competitive read mapping. a) Dynamics of NS5 genera based on representative type species and b) dynamics of species within a single genus, *Candidatus* Marivariicella.

between 20 – 46°C [98]. In the marine environment, temperature variations occur over depth and latitudinal gradients, ranging from >30°C in the equatorial Atlantic to <0°C in polar waters. Along these gradients, the optimal temperature ranges of different microbes will occur. This is supported by the frequent observation that temperate correlates strongly with microbial dynamics over spatial scales in the ocean [93, 94]. Similarly, in chapter II of this thesis, we observed inter- and intra-genus differences in the dynamics of NS5 members with temperature over a globally distributed dataset (Figure 4). Large variations were also observed in the breadth of temperatures under which species reached peaked relative abundance values, indicating narrow and broad adaptations. For small photosynthetic prokaryotes of the *Prochlorococcus*, a similar observation has been made at the sub-species level [93]. Collectively, these findings highlight the importance of temperature for defining ecological niche space over spatial scales.

6.3.1.2 Oxygen

Despite microbial life beginning under anoxic conditions, the oceans today are predominantly oxygenated and thus support aerobic microbial respiration. It is estimated that oxygen fuels up to 90% of microbial organic matter respiration in the oceans [99]. With that, it could be expected that oxygen would play a limited role in shaping ecological niche space within the oceans. This is, in part, what was observed in this thesis. Although correlations with NS5 species abundance and oxygen concentrations over large spatial scales were observed, it is likely that these patterns were reflective of dynamics in relation to water masses. For example, polar waters in the Fram Strait are known to be more oxygenated than inflowing Atlantic waters (Chapter IV). However, there are a number of considerations warranted for oxygen with respect to ecological niche space in pelagic marine environments. Of particular significance is the expansion of pelagic areas where oxygen is severely depleted, termed oxygen minimum zones (OMZs). The periphery of OMZs is characterised by oxygen gradients that favour a succession of aerobic to anaerobic microbial processes [100], reflecting the redox-driven partitioning of microbes in sediments and marine snow [101]. In contrast, the core of OMZs is dominated by anaerobic microbial processes. It is thus generally assumed that OMZs would act as a barrier to aerobic microbes. However, OMZs do not necessarily restrict photosynthesis, with deep chlorophyll maxima being evidenced in association with and within OMZs, with the subsequent oxygen production capable of supporting microbial aerobic respiration [102, 103]. Physical processes can also provide ventilation to the oxycline in the upper OMZ regions, providing pulses of oxygen-rich water [104]. A number of other findings are also redefining our understanding on microbial respiration of pelagic marine microbes and are important to consider with respect to ecological niche space. Firstly, is the observation that some microbes previously considered to be aerobic, actually harbour the capacity to use

nitrate as an electron acceptor [105]. Furthermore, the oxygen threshold of microbial aerobic respiration is constantly being lowered [106] and in sediments, microbes have been shown to simultaneously perform aerobic and anaerobic respiration [107]. These findings suggest that the availability of oxygen has a heterogeneous effect across the microbial tree of life and our preconceived boundaries for microbial respiration may not be so clearly reflected in the environment. Concurrent with the expansion of OMZs [108] is the deoxygenation of the global oceans [109, 110], which combined could increase the importance of oxygen for shaping the ecological niches of pelagic microbes in the future.

6.3.1.3 Irradiance and daylight length

The availability of light is the primary determinant of photosynthesis and thus primary production in the upper marine water column. As microbes are reliant on the organic matter synthesised by primary producers, light has a major indirect role on microbial growth and dynamics. However, it can also exert a direct influence. In the upper marine water column, it is estimated that ~50% of microbes harbour rhodopsins [111], membrane-bound light-driven ion pumps. The generation of biochemical energy from rhodopsins can not only enhance the fitness of microbes but also promote survival during starvation [112]. Microbial rhodopsins are also spectrally tuned to absorb different wavelengths of light, which determines their efficiency over vertical and latitudinal gradients. For example, it has been shown that blue light-absorbing rhodopsins are more abundant in sub-surface waters [113]. In Chapter III of this thesis, we showed that all NS5 Marine Group species harboured green light-absorbing proteorhodopsins and were also more prevalent in surface waters. This suggests that for rhodopsin-containing microbes, irradiance could play a role in determining spatial dynamics over lateral and vertical scales and thus contribute to defining niche space.

In Chapter IV of this thesis, we identified that daylight length is a key factor shaping microbial community composition in polar and Atlantic water masses. However, upon deeper investigation, the observed changes are more likely a direct influence of primary production and organic matter availability.

6.3.1.4 Organic substrate availability

Although marine microbes exhibit a huge diversity of lifestyles and metabolic capacities, they are all reliant on exogenous organic compounds for growth. However, the extent of this reliance is highly heterogeneous across phylogenetic groups and reflects adaptations to different environments, natural selection and genetic drift. Although some microbes can fix their own carbon and synthesise the majority of essential building blocks, many pelagic marine microbes cannot. This has been proposed as a reductive evolutionary process that drives dependency through gene loss, also known as the “Black Queen” hypothesis [114]. By

minimising metabolic cost, a higher growth efficiency can be achieved. Regardless of the microbe in question, it is thus expected that the availability of the required organic substrate is likely to be a key factor that contributes to defining ecological niche space.

The importance of organic substrate for influencing microbial niche space has been well evidenced in relation to spring phytoplankton blooms in coastal temperate ecosystems [44]. These phenomena result in the release of large amounts of organic matter that fuel the growth of heterotrophic microbes. Multi-year studies have evidenced a recurrent, deterministic response of some microbial clades, with a successional like pattern being attributed to substrate-based niche partitioning [43, 115]. In Chapter IV, we identified that heterotrophic microbial clades functionally linked to phytoplankton-derived organic matter also experience pronounced, ephemeral dynamics under polar-day and Atlantic water mass conditions. Although in that study we lacked consistent chlorophyll *a* measurements, we believe that the observed dynamics would closely follow phytoplankton blooms. In addition, under polar night and high-ice conditions, the community was dominated by organisms that either harboured autotrophic capacities or were enriched in genes for the degradation of bacterial- and terrestrial-derived organic matter and more recalcitrant compounds that likely persisted from the productive season. This indicates that substrate availability also shapes temporal dynamics of microbes and contributes to shaping niche space in high-latitude waters.

Using carbohydrates as a focal group of organic compounds, we showed in Chapter V that substrate availability and substrate utilisation by microbial populations is highly heterogeneous over spatial scales. The monosaccharide and polysaccharide fractions of particulate organic matter were evidenced to change in composition and abundance over depth and lateral spatial scales in high North Atlantic waters at the end of summer. Concurrently, the distribution of microbial populations and the transcription of genes related to carbohydrate utilisation exhibited similar heterogeneous patterns. As Chapter V represents one of the first to perform a high-resolution assessment of substrate availability in conjunction with microbial substrate utilisation over spatial scales, several important observations were made that contribute to our understanding of substrate on defining ecological niche space. Firstly, populations with metabolic capacities to use diverse substrates for growth are able to change transcription patterns over spatial scales to target different substrates. Secondly, within each sample, some substrates were targeted by many populations, such as β-1,3-glucan, while others were only targeted by a single population, such as alginate, which indicates substrate-based niche partitioning. This mirrors the findings observed in temporal-based studies [44, 116, 117] and highlights that organic substrate availability plays an important role in defining ecological niche space over space and time.

Our current understanding on the role substrate plays in defining ecological niches over space and time is still in its infancy and further investigations are needed. However, there

are a number of challenges with this. Typically, to gain accurate identification of compounds, specialised techniques are required, which are limited by the standards and reference data available. For example, the carbohydrate-microarray approach used in Chapter V to identify polysaccharide epitopes is limited by the number of antibodies available. For more simple substrates that are widely used by bacteria, their rapid turnover times can result in consistently low concentrations that are undetectable by analytical methods. In the case of dissolved organic compounds, usually hundreds to thousands of litres of water must be concentrated down to reach detectable quantities, which is timely and unfeasible in many projects. Lastly, in order for us to identify the organic substrates used by environmental populations, we rely on annotating genes against databases that in themselves are limited to enzymes that have been biochemically analysed.

These are only some of the most pressing difficulties with investigating substrate utilisation by microbes and the impact of substrate availability on niche space. However, continued developments in technologies and software are opening up promising new avenues to alleviate some of the above-raised limitations. Of particular note, is the recent introduction of AlphaFold [118]. AlphaFold is an AI-based algorithm that can predict protein structure from an amino acid sequence with an incredible degree of accuracy. By knowing the structure of a protein, its catalytic reaction and target substrate can be determined. This could lead to a tremendous increase in protein characterisation that can expand databases and subsequently reduce the proportion of 'unknown' genes in environmentally derived microbial genomes. Concurrent with this, a more widespread application of metatranscriptomics and metaproteomics could elucidate patterns in microbial substrate utilisation over spatiotemporal scales.

6.3.1.5 Other factors

There are countless other biotic and abiotic conditions that have a varying degree of influence on shaping the ecological niche space of marine microbes over space and time. Some of which include viral activity, predation/grazing, mutualistic interactions and inorganic compound availability (e.g. nitrate). Within the scope of this discussion, a consideration of all conditions is not warranted. There is however, one noteworthy component that was discussed within the work of this thesis and is often overlooked. That is, physical hydrographic processes. In Chapter IV, we observed that water mass was the single most important factor correlating to changes in microbial community composition, which is in agreement with previous studies [53, 54, 119, 120]. Although water mass in itself, is not an informative factor, due to it representing a vast number of different physicochemical conditions, it does highlight that it is important to consider physical oceanography in microbial distribution patterns and ecological niche space.

As members of the plankton, physical processes can influence microbes across all spatial scales, from determining the contact rate of non-motile bacterium with their substrate to involuntary large-scale dispersal through oceanic currents. In addition to this, differences in physicochemical properties of water masses act as strong physical boundaries that can limit dispersal of microbes. Although such process may not be considered as niche-defining factors, they are of paramount importance in understanding the spatial and temporal restrictions of niches, the emergence of niches and whether microbes can respond to them.

6.3.2 Redefining the ecological niche concept for marine microbial ecology

The niche concept provides a testable prediction and a foundation for research into the ecology of microbes, which in itself reveals valuable information on evolution, diversification and how biotic and abiotic conditions shape microbial dynamics. However, the adoption of the Hutchinson defined niche concept in microbial ecology without any alterations has led to varied interpretations and its application in different contexts and at different phylogenetic resolutions. Therefore, in order to rectify inconsistency in its usage and to lay a clear foundation for future research, it is necessary to modify and update the concept in the context of microbial ecology. The most important components to reconsider from the original concept are the idea of conditions shaping the niche within an environment and those conditions defining the niche of a species. In particular for marine microbes, environmental boundaries are not easily discernible and similar conditions can arise in different geographical locations. Furthermore, do niches partition at the level of species for microbes?

Extensive efforts have resulted in phylogenetic boundaries for microbial species being operationally defined, with a threshold of 95% average nucleotide identity (ANI) [121, 122]. Although there can be deviations, this ANI threshold generally holds across the available genomes in public databases. However, it is now well evidenced that there exists a high degree of microdiversity in microbial species, that is, genetic variations at the sub-species level [123, 124]. Although terms to describe sub-species variants have been devised, such as strain or phylotype, do these represent ecologically distinct units or not? If so, is this the phylogenetic level upon which niches partition? I believe that niches do partition at the sub-species level, which I will discuss in more detail after introducing an important, more recent model/concept.

6.3.2.1 The ecotype model

Frederick Cohan introduced the ecotype model in 2006 [125] to try to address the partitioning of microbial niches An ecotype was defined as, “a group of bacteria that are ecologically similar to one another”, which show a history of coexistence, as inferred by phylogeny, and show a prognosis for future coexistence based on ecological distinctness between ecotypes. In that

model, a single ecotype occupies a single ecological niche. Genetic diversity within an ecotype can occur but only until a periodic selection event, where it is purged and the most successfully genetic variant emerges. If genetic variations accumulate to an extent where a new ecological niche can be occupied, then a new ecotype emerges. Cohan also stated that although ecotypes are ecological units, they can be partitioned based on sequence phylogeny.

In its foundation, I believe the model has merit and it harbours many of the necessary elements for a clear niche concept to be defined in the context of microbial ecology. However, there are other aspects about the ecotype model which need further consideration. Most notably, is that the exact definition of an ecotype is too broad and generalised and, as with the adoption of the niche concept in microbial ecology, is open to interpretation and varied applications. In addition, I believe that the model was over simplified, with no consideration for overlapping niche space. Lastly, the model lacked more thorough extrapolations to environmental settings, and completely lacked considerations of marine microbes.

It is with the considerations above that I propose a new model and ecological niche concept, which provides clearer definitions and may be more widely applicable in microbial ecology, particularly in the marine environment.

6.3.2.2 The modified ecological niche concept

Investigations into microbial growth in cultivation-based research has long shown that microbes are able to grow under a wide range of conditions, but typically reach optimal growth under a more narrow range. This is particularly well evidenced for temperature. A study investigating growth dynamics of marine SAR92 Clade isolates revealed several that could grow between 4 – 25°C, but exhibited optimal growth between 16 – 20°C [126]. In the same study, they showed variable growth dynamics in relation to the addition of dissolved organic carbon in the media. SAR92 isolates exhibited stable growth dynamics under no DOC addition and up to a fivefold increase in DOC, but subsequently collapsed under tenfold increases in DOC concentration. Another important consideration in addition to ranges of conditions, is with respect to organic substrates. Many heterotrophic microbes are capable of growth on a range of organic carbon substrates, which was also reflected in Chapter IV of this thesis through variations in the transcription of carbohydrate degrading genes by a single MAG over sampling stations. However, the energy and element yield from each substrate is heterogeneous and thus, one substrate may facilitate higher growth rate and sustain larger population sizes than another.

To encapsulate these ecological observations, I propose a demarcation of niches into “optimal” and “sub-optimal”. I believe that this not only better reflects the dynamics of natural microbial populations but also allows for a clearer understanding of how diversification to fill new niches occurs.

Optimal niche:

"The optimal niche describes a set of conditions, irrespective of geographical location, that allow for optimal growth of an ecovariant"

Sub-optimal niche:

"The sub-optimal niche describes a set of conditions, irrespective of geographical location, that allow for a degree of growth of an ecovariant"

Along with the introduction of two niche definitions, a new term should also be introduced, ecovariant. The term ecovariant is but a modification of Cohan's ecotype, but will be defined more clearly.

Ecovariant:

"A sub-species variant that is genetically and ecologically distinguishable"

There are two components to the ecovariant that require elaboration. In terms of genetic discernibility, the key is that an ecovariant harbours a distinct genetic backbone to other ecovariants. This backbone is maintained among all members of that ecovariant and is essential for occupying the ecological niche however, genetic variations can exist around that backbone. In agreement with Cohan's ecotype model, when the genetic variation provides a capacity to occupy a new niche, then a new ecovariant emerges. The other component of the ecovariant definition is in terms of ecological distinction. Of course, this refers to its occupation of a distinct niche, but there other routes that could be taken in order to ecologically distinguish variants without the need to describe their complete niche. I think the most valuable aspect here is spatiotemporal dynamics. Given the wealth of genetic data available that has been generated from all over the world's oceans (spatial) and over multiple years in fixed locations (temporal), the dynamics of variants can be determined. As the niche describes a set of conditions that occur in a given space and time, two ecovariants should be distinguishable based on their spatiotemporal dynamics. With this, we can infer niches *a priori* to performing more detailed genomic analysis.

6.3.2.3 Example of predicting ecovariants based on spatiotemporal dynamics

Using data generated throughout the work of this thesis, I try to provide an example of how we could discern ecovariants using spatiotemporal dynamics. The example includes eleven MAGs (Table 1) that belong to the *Candidatus* Maricapacicella, one of the candidate genera designated in Chapter III, that share >98.5% average nucleotide identity (ANI) (Figure 5).

Comparing the dynamics across PacBio HiFi read metagenomes from the Fram Strait (generated during this thesis), resulted in two large clusters. Within each of those clusters the MAGs could be further distinguished, resulting in four potential ecovariants, termed ecoA – ecoD. Based on these dynamics, we could hypothesise that the optimal niche of ecoA, is the sub-optimal niche of ecoB, whilst the optimal niche of ecoD, is the sub-optimal niche of ecoC but the niches of ecoA+ecoB do not overlap with those of ecoC+ecoD. Another possibility is that these clusters represent two ecovariants (ecoAB and ecoCD), with the MAGs comprising genetic variants within the ecovariants that contribute varying proportions to the abundance of the whole ecovariant in each sample. These observations provide the basis of ecological distinction between the ecovariants, but what about genetic distinction? The ANI values within and between these preliminary ecovariants were comparable, suggesting that, at least for this example, ANI may not be a useful tool for discerning ecovariants. However, ANI is only calculated based on shared genetic content between two genomes and thus reflects single nucleotide variations but fails to capture gene duplication/loss/gain. Although this is but one preliminary example that needs further investigation, it provides the basis for how ecovariants may be distinguished. I believe that ecovariants are most easily discernible by distinct distributions over spatiotemporal scales, which would provide a foundation for investigations at the genomic level.

Table 1. Summary statistics for 11 MAGs obtained during this thesis that are used as an example for ecovariant demarcation. The MAGs were derived from the PacBio HiFi metagenomes generated during Chapter I. Their summary statistics were obtained using CheckM.

MAG	Comp. (%)	Cont. (%)	Size (Mbp)	Number of contigs	N50 (bp)	GC content
FRAM18_4502_E_metabat_bin_09	79.41	0	1.295	108	15087	33.37
FRAM18_4502_F_metabat_bin_06	67.8	0	1.102	90	17558	33.48
FRAM18_4502_G_metabat_bin_08	88.24	0	1.335	87	25808	33.49
FRAM18_4502_H_metabat_bin_09	68.05	1.9	1.207	179	7970	33.57
FRAM18_4502_I_concoct_bin_14	77.52	0	1.462	150	16017	33.25
FRAM18_4502_J_concoct_bin_07	87.87	0	1.674	67	41251	33.4
FRAM18_4514_A_concoct_bin_09	91.54	0.03	1.614	70	35980	33.42
FRAM18_4514_B_concoct_bin_16	88.6	0.8	1.558	62	33828	33.55
FRAM18_4571_PB_H_metabat_bin_09	87.13	0	1.606	10	200090	33.61
FRAM18_4571_PB_H_metabat_bin_10	91.54	0	1.829	12	227436	33.55
FRAM18_4571_PB_I_metabat_bin_01	91.54	0	1.859	11	221399	33.54

6.3.2.4 Are there spatial boundaries within which niche conditions are defined?

The original niche concept stipulated that conditions within an environment define the niche of a species in that environment. Indifference to many terrestrial habitats, environmental boundaries are not so easily discernible in the marine environment. In order to be more widely applicable in marine microbial ecology, I thus excluded the environment from the modified

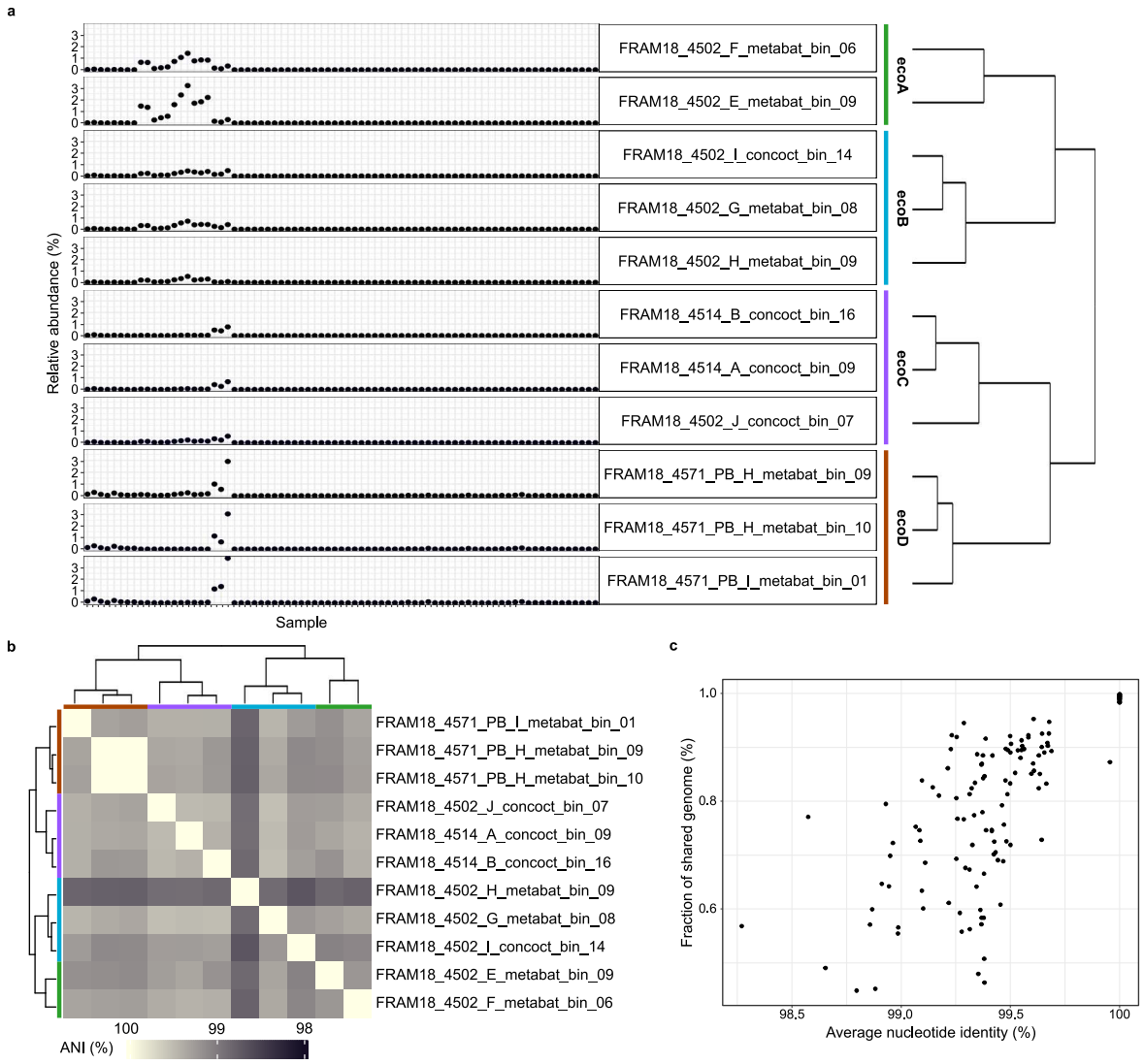


Figure 5. Distribution dynamics and average nucleotide identity comparisons between the eleven MAGs from the selected species. A single species, assigned to the Candidatus Maricapacicella, which contained eleven MAGs that shared a average nucleotide identity (ANI) of >98.5% were selected. A) The relative abundance of MAGs was determined across the Fram Strait metagenomes generated during this thesis. Relative abundance is defined as the quotient between the truncated average depth (TAD80) and the number of genomes captured in each metagenome. The dendrogram was determined based on a dissimilarity matrix computed from the relative abundances. B) ANI values between all MAGs ordered based on the dendrogram from part a). c) Shared ANI values between all MAGs in comparison to fraction of shared genome, from which ANI was calculated.

definition. Despite this, I do believe there are many cases where the spatial extent of a niche could be defined in a form of environmental boundary. An extreme example would be fluids of hydrothermal vents that are geochemically distinct from the surrounding seawater and harbour unique microbial assemblages [127, 128]. The more challenging case, is for marine pelagic microbes. However, in the pelagic realm, changes in conditions are typically observed across the boundaries of water masses. Water masses are bodies of water with homogeneous physicochemical conditions. As microbes are involuntarily transported with water movements, they are, to some degree, bound by water masses. As such, water masses could define the geographical spatial extent within which ecovariants can respond to their niches. To put this into perspective, we can consider time-series analyses. Time-series studies have been able to track the dynamics of microbial clades at the scale of days, weeks and months [43, 44, 84, 129]. For example, at Helgoland Roads time-series, it has been shown that microbial clades are able to rapidly proliferate over the order of days in response to spring phytoplankton blooms [43, 44]. To track such events, water samples are collected at multi-day intervals. In the region around Helgoland, current velocities are typically around 0.5 m^2 s^{-1} [130]. As such, even with sampling at multiday intervals, the actual body of water that has been sampled over that time frame is potentially on the order of tens of km^2. Therefore, if the dynamics of microbial clades are able to be tracked over the samples, it indicates that the spatial extent of the ecological niche of those clades is large. I believe the limits of that area, are the boundaries of the water mass. Considering the complex hydrographic dynamics of the oceans, such boundaries are not always operationally easy to define and there are numerous physical processes that result in the mixing of water masses and thus the potential dispersal of microbes. Nonetheless, I believe where a water mass can be identified, its boundaries would define the potential spatial extent of the ecovariant's niche at any given time.

6.3.2.5 How do ecovariants respond to niches and how do new ecovariants emerge?

In order for an ecovariant to respond to its niche it must be present in the water body. As members of the plankton, microbes are involuntarily transported with water movements and to date, no evidence has been shown to suggest microbes can retain themselves within a location. As such, some fraction of the ecovariant population must be able to persist through unfavourable conditions in order to act as seeds [131]. This has been proposed as the microbial seed bank hypothesis [132]. Although it was largely formulated to describe the seeding of populations within an environment, I propose that it is valid for the wider marine environment. The hypothesis stipulates that microbes can enter a reversible state of dormancy (highly reduced metabolic activity) to persist through unfavourable conditions and later rejuvenate in response to changes in the environment. This theory provides the only logical explanation as to how an ecovariant could appear and respond to its niche. If we accept this

hypothesis into our modified niche concept, it also would agree with “everything is everywhere, but the environment selects” that was proposed by Baas-Becking nearly 100 years ago [133]. Furthermore, the ability of microbes to enter dormancy would provide an explanation and answer to the “paradox of the plankton”. This paradox was raised by Hutchinson in 1961 [134] to question how so many planktonic species can coexist in the same environment given a limited number of resources – originally proposed for phytoplankton but it also has been discussed in the context of microbes.

In order for a new ecovariant to emerge, sufficient genetic modifications must have taken place that allow it to occupy a new niche. As the ecovariant is likely in a dormant and non-replicating state for long periods of time outside of its niche conditions, I believe that the necessary mutation and recombination events could only take place within optimal niche or sub-optimal niche conditions. Furthermore, I posit that due to evolutionary selective pressure, genetic modifications under sub-optimal niche conditions would more likely result in the emergence of a new ecovariant. As under sub-optimal growth conditions, any genetic modifications that result in improved fitness will likely be sustained. Over evolutionary time-scales, multiple modifications could take place that now turn those sub-optimal conditions into the optimal conditions for that particular genetic variant, forming a new ecovariant. A very oversimplified, conceptual schematic is provided to illustrate this (Figure 6). If an ecovariant continues to acquire genetic modifications through periodic selection or other evolutionary processes, then it will likely result in a new species emerging, thus ecovariants could be considered precursors to speciation.

6.3.2.6 Conceptualising niche space of ecovariants in the environment

As an ecovariant must emerge from a pre-existing ecovariant, it is within reason to assume that both will be able to grow under similar conditions. I believe that in the marine environment, there is much overlap in sub-optimal niche space between ecovariants. Sub-optimal niche space likely overlaps for multiple ecovariants within a species and the sub-optimal niche space of one ecovariant likely represents the optimal niche space for another ecovariant, as shown in the example above. The only clear distinction is that no two ecovariants will have overlapping optimal niches.

As an example, we can consider the extensive work carried out on marine *Prochlorococcus*. Although the phylogeny of this lineage is still under debate and future studies may prove to change the current perspective, it still provides an example for consideration. *Prochlorococcus* are one of the most numerically abundant microbes and dominate the photosynthetic biomass in oligotrophic regions of the oceans. Using single-cell genomics, Kashtan et al. [135] identified numerous sub-species level genetic variants of *Prochlorococcus* and showed they exhibited distinct distribution patterns over temporal scales.

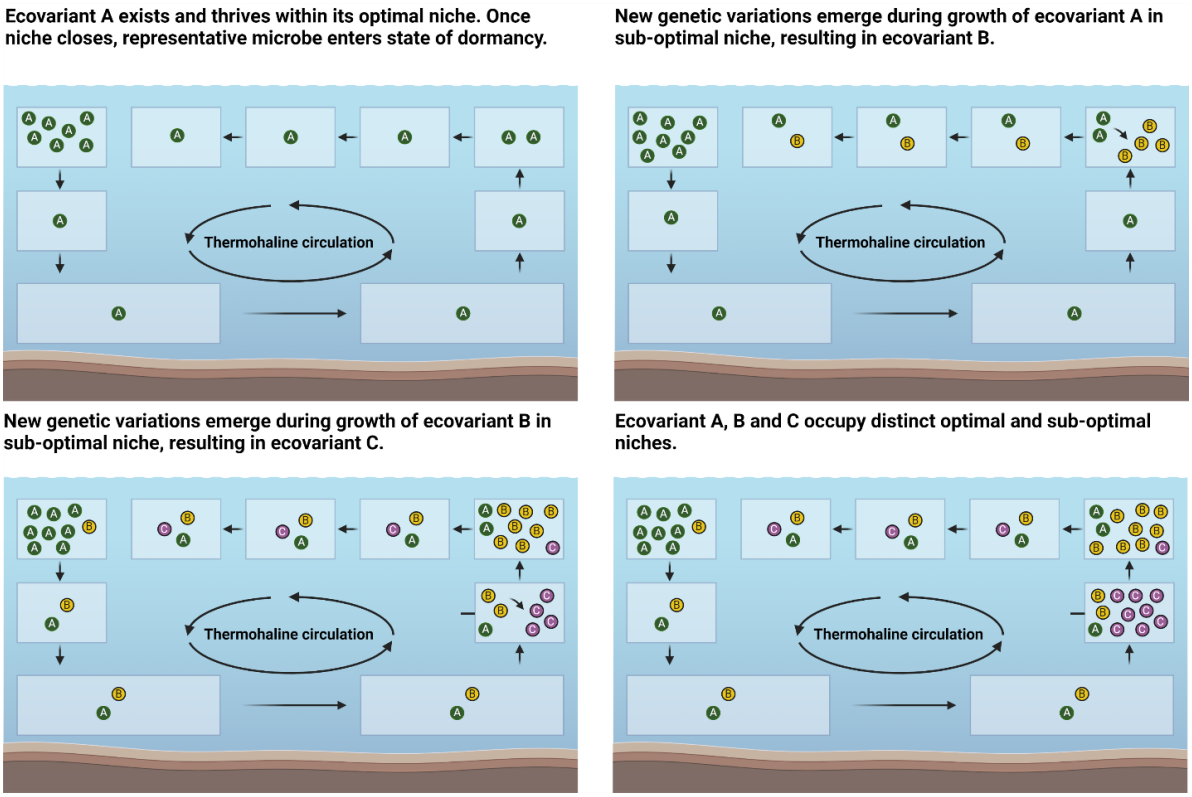


Figure 6. Conceptual diagram illustrating the evolution of ecotypes and their growth under optimal and sub-optimal conditions. A simplified view of the oceans is shown with the boxes representing different conditions. Through the thermohaline circulation, ecotypes are transported through the world's oceans over long time periods (order of hundreds to thousands of years), and pass from optimal to sub-optimal niches that allow for growth separated by unfavourable conditions where the microbe enters a state of dormancy or highly reduced metabolic activity.

The genetic differences between these variants were attributed to single-nucleotide polymorphisms and a small number of gene content differences. However, each of the variants harboured a stable, distinguishable genomic backbone. They further evidenced that for some of the variants, distinct temporal dynamics could be related to changes in conditions, indicative of optimal niche partitioning. However, all variants also overlapped, which could reflect overlapping sub-optimal niches as proposed above. There were limitations in that study, pointed out also by the authors themselves, particularly with respect to only 70% of the genome for each sub-species variant being acquired. Whether or not these will prove to represent distinct ecovariants is yet to be determined however, the general observations are in agreement with the concept presented here.

6.3.2.7 Genomically discerning ecovariants

As mentioned previously, ecovariants describe a distinct genetic variation within a species but it does not mean that all microbes within an ecovariant are genetically identical. The question that then arises is, can we genetically distinguish ecovariants? And if so, on what basis? As was preliminarily shown in the study by Kashtan et al. outlined above, I believe that ecotypes

will be genetically distinguishable. My reasoning is that, an ecovariant has acquired some genomic modifications that allow it to occupy a distinct niche and thus all microbes within that ecovariant must harbour those features to continue occupying that optimal niche. As such, an ecovariant must have a core genomic signature that can be identified, whether this be in the form of cohesive single nucleotide polymorphism profiles or gene content or a combination of both. A recent study, currently only available on BioRxiv, has suggested that sub-species variations in the form of strains or clonal complexes can be defined by a 99.5% ANI cut-off threshold [136]. However, those findings were based only on cultured isolate genomes and I believe that in the natural environment, such a definitive threshold may not hold. In the example I provided above using MAGs generated during this thesis, we observed distinct dynamics for MAGs that share >99.5% ANI, which would oppose such a definitive threshold (Figure 5). However, to better understand the genetic variation between ecovariants, more extensive analysis using high-quality metagenome-assembled genomes is required.

6.3.3 Outlook on ecological niches and the proposed modified concept

6.3.3.1 Improving our understanding on ecological niches in the future

The rapid development of new technologies, the ever-increasing resolution and scale of environmental sampling campaigns and the advancements in computational software and hardware are continually opening up new avenues for investigations on the ecology of microbes in the marine environment. Over the past ten years, global-scale sampling campaigns, such as those conducted by Tara Oceans [137] and Malaspina [138], in conjunction with numerous more localised, higher resolution sampling efforts has greatly expanded the spatial coverage of microbial genomic data in the oceans. In addition, the extension of time-series campaigns and the regular sampling at long-term ecological research stations is providing an ever-increasing temporal coverage of microbial genomic information within environments. These datasets have revolutionised our understanding on marine microbial ecology. However, genomic data alone is insufficient for understanding ecological niches. As ecological niches define a set of conditions, it is essential to investigate genomic content and spatiotemporal dynamics of microbes in the context of oceanographic, chemical and climate data as well as observations on other trophic levels, such as primary producers. A plethora of research on these other environmental components in recent years has also lead to an expansion in such data. Moving forward, it is going to be essential that ecologists actively explore the data and findings from these other scientific disciplines if we hope to deepen our understanding of ecological niches. Alternatively, and the significantly better option, researchers from different disciplines should combine efforts in an ecosystem-targeted approach, whereby all biotic and abiotic components are analysed simultaneously.

6.3.3.2 Final considerations on the modified niche concept

It is important to state that the proposed modifications to the ecological niche concept outlined above are a “working theory”. The aim of redefining the concept is to place it into the context of our current understanding on microbial ecology whilst connecting it to more recently proposed theories and hypotheses. In addition, by redefining the niche concept, I hope to inspire further research that aims to prove or disprove it, which will only lead to continued developments in our understanding of ecology along the way. During the course of writing this thesis, the concept evolved greatly and I continue to reconsider its components by placing it into different scenarios to ascertain whether it could hold or not. With that, I would like to finish by providing some additional “food for thought” and considerations on the modified concept:

- It may be that an optimal ecological niche for a microbe only exists once, in a given space and time.
- There may be genetically distinct microbial populations capable of occupying the same optimal niche and stochastic processes determine the winner in a given space and time.
- It is possible that for many microbes, we may not be able to identify or discern their optimal and sub-optimal niches. This can occur if the optimal niche is ephemeral and is simply not captured during sampling efforts. It is likely also the case for microbes that only exist in the rare biosphere.

References

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

2. Arrigo KR. Marine microorganisms and global nutrient cycles. *Nature*. 2005;437:349–355.

3. Field CB, Raupach MR. The global carbon cycle: integrating humans, climate, and the natural world. 2004. Island Press.

4. Toggweiler JR, Key RM. Thermohaline circulation. In: Encyclopedia of Ocean Sciences, eds (J. H. Steele). 2001. Academic Press, pp 2941–2947.

5. Bigg GR, Jickells TD, Liss PS, Osborn TJ. The role of the oceans in climate. *Int J Climatol*. 2003;23:1127–1159.

6. Herman Y. Topography of the Arctic Ocean. In: Marine Geology and Oceanography of the Arctic Seas, eds (Herman. Y.). 1974. Springer, Berlin, Heidelberg, pp 73–81.

7. McClelland JW, Holmes RM, Dunton KH, Macdonald RW. The Arctic Ocean estuary. *Estuaries and Coasts*. 2012;35:353–368.

8. Assmy P, Ehn JK, Fernández-Méndez M, Hop H, Katlein C, Sundfjord A, et al. Floating ice-algal aggregates below melting Arctic sea ice. *PLOS ONE*. 2013;8:e76599.

9. Kauko HM, Taskjelle T, Assmy P, Pavlov AK, Mundy CJ, Duarte P, et al. Windows in Arctic sea ice: light transmission and ice algae in a refrozen lead. *J Geophys Res Biogeo*. 2017;122:1486–1505.

10. Kwok R. Arctic sea ice thickness, volume, and multiyear ice coverage: losses and coupled variability (1958–2018). *Environ Res Lett*. 2018;13:105005.

11. Notz D, Community S. Arctic sea ice in CMIP6. *Geophys Res Lett*. 2020;47:e2019GL086749.

12. Årthun M, Eldevik T, Smedsrud LH, Skagseth Ø, Ingvaldsen RB. Quantifying the influence of Atlantic heat on Barents sea ice variability and retreat. *J Clim*. 2012;25:4736–4743.

13. Oziel L, Baudena A, Ardyna M, Massicotte P, Randelhoff A, Sallée J-B, et al. Faster Atlantic currents drive poleward expansion of temperate phytoplankton in the Arctic Ocean. *Nat Commun*. 2020;11:1–8.

14. Neukermans G, Oziel L, Babin M. Increased intrusion of warming Atlantic water leads to rapid expansion of temperate phytoplankton in the Arctic. *Global Change Biol*. 2018;24:2545–2553.

15. Serreze MC, Barrett AP, Slater AG, Woodgate RA, Aagaard K, Lammers RB, et al. The large-scale freshwater cycle of the Arctic. *J Geophys Res Oceans*. 2006;111:C11010.

16. de Steur L, Hansen E, Mauritzen C, Beszczynska-Möller A, Fahrbach E. Impact of recirculation on the East Greenland Current in Fram Strait: results from moored current meter measurements between 1997 and 2009. *Deep Sea Res Part I: Oceanogr Res Pap*. 2014;92:26–40.

17. Hofmann Z, von Appen W-J, Wekerle C. Seasonal and mesoscale variability of the two Atlantic water recirculation pathways in Fram Strait. *J Geophys Res Oceans*. 2021;126:e2020JC017057.

18. Sanger F, Air GM, Barrell BG, Brown NL, Coulson AR, Fiddes JC, et al. Nucleotide sequence of bacteriophage φX174 DNA. *Nature*. 1977;265:687–695.

19. Woese CR, Fox GE. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. *Proc Natl Acad Sci*. 1977;74:5088–5090.

20. Fox GE, Pechman KR, Woese CR. Comparative cataloging of 16S ribosomal ribonucleic acid: molecular aproach to procaryotic systematics. *Int J Syst Evol Microbiol*. 1977;27:44–57.

21. Sogin ML, Morrison HG, Huber JA, Welch DM, Huse SM, Neal PR, et al. Microbial diversity in the deep sea and the underexplored “rare biosphere”. *Proc Natl Acad Sci*. 2006;103:12115–12120.

22. Eren AM, Esen ÖC, Quince C, Vineis JH, Morrison HG, Sogin ML, et al. Anvi’o: an advanced analysis and visualization platform for ‘omics data. *PeerJ*. 2015;3:e1319.

23. Nishimura Y, Yoshizawa S. The OceanDNA MAG catalog contains over 50,000 prokaryotic genomes originated from various marine environments. *Sci Data*. 2022;9:305.

24. Vernette C, Lecubin J, Sánchez P, Tara Oceans Coordinators, Sunagawa S, Delmont TO, et al. The ocean gene atlas v2.0: online exploration of the biogeography and phylogeny of plankton genes. *Nucleic Acids Res*. 2022;50:W516–W526.

25. Tuerena RE, Hopkins J, Buchanan PJ, Ganeshram RS, Norman L, von Appen W-J, et al. An Arctic strait of two halves: The changing dynamics of nutrient uptake and limitation across the Fram Strait. *Global Biogeochem Cycles*. 2021;35:e2021GB006961.

26. Wietz M, Bienhold C, Metfies K, Torres-Valdés S, von Appen W-J, Salter I, et al. The polar night shift: seasonal dynamics and drivers of Arctic Ocean microbiomes revealed by autonomous sampling. *ISME Commun*. 2021;1:1–12.

27. Reigstad M, Carroll J, Slagstad D, Ellingsen I, Wassmann P. Intra-regional comparison of productivity, carbon flux and ecosystem composition within the northern Barents Sea. *Prog Oceanogr*. 2011;90:33–46.

28. Cherkasheva A, Bracher A, Melsheimer C, Köberle C, Gerdes R, Nöthig E-M, et al. Influence of the physical environment on polar phytoplankton blooms: A case study in the Fram Strait. *J Mar Syst*. 2014;132:196–207.

29. Gradinger RR, Baumann MEM. Distribution of phytoplankton communities in relation to the large-scale hydrographical regime in the Fram Strait. *Mar Biol*. 1991;111:311–321.

30. Amon RMW, Benner R. Combined neutral sugars as indicators of the diagenetic state of dissolved organic matter in the Arctic Ocean. *Deep Sea Res Part I: Oceanogr Res Pap*. 2003;50:151–169.

31. Anderson LG, Olsson K, Skoog A. Distribution of dissolved inorganic and organic carbon in the Eurasian Basin of the Arctic Ocean. *Geophys Monogr Ser*. 1994;85:255–262.

32. Amon RMW, Budéus G, Meon B. Dissolved organic carbon distribution and origin in the Nordic Seas: Exchanges with the Arctic Ocean and the North Atlantic. *J Geophys Res*. 2003;108:3221.

33. Pavlov AK, Granskog MA, Stedmon CA, Ivanov BV, Hudson SR, Falk-Petersen S. Contrasting optical properties of surface waters across the Fram Strait and its potential biological implications. *J Mar Syst*. 2015;143:62–72.

34. Granskog MA, Stedmon CA, Dodd PA, Amon RMW, Pavlov AK, Steur L de, et al. Characteristics of colored dissolved organic matter (CDOM) in the Arctic outflow in the Fram Strait: Assessing the changes and fate of terrigenous CDOM in the Arctic Ocean. *J Geophys Res Oceans*. 2012;117.

35. Opsahl S, Benner R, Amon RMW. Major flux of terrigenous dissolved organic matter through the Arctic Ocean. *Limnol Oceanogr*. 1999;44:2017–2023.

36. von Jackowski A, Grosse J, Nöthig E-M, Engel A. Dynamics of organic matter and bacterial activity in the Fram Strait during summer and autumn. *Philos trans, Math phys eng sci*. 2020;378:20190366.

37. Engel A, Bracher A, Dinter T, Endres S, Grosse J, Metfies K, et al. Inter-annual variability of organic carbon concentration in the eastern Fram Strait during summer (2009–2017). *Front Mar Sci*. 2019;6.

38. Nöthig E-M, Ramondenc S, Haas A, Hehemann L, Walter A, Bracher A, et al. Summertime chlorophyll a and particulate organic carbon standing stocks in surface waters of the Fram Strait and the Arctic Ocean (1991–2015). *Front Mar Sci*. 2020;7.

39. von Jackowski A, Becker KW, Wietz M, Bienhold C, Zäncker B, Nöthig E-M, et al. Variations of microbial communities and substrate regimes in the eastern Fram Strait between summer and fall. *Environ Microbiol*. 2022;24:4124–4136.

40. Vidal-Melgosa S, Sichert A, Francis TB, Bartosik D, Niggemann J, Wichels A, et al. Diatom fucan polysaccharide precipitates carbon during algal blooms. *Nat Commun*. 2021;12:1150.

41. Fadeev E, Salter I, Schourup-Kristensen V, Nöthig E-M, Metfies K, Engel A, et al. Microbial communities in the east and west Fram Strait during sea ice melting season. *Front Mar Sci*. 2018;5.

42. Gilbert JA, Field D, Swift P, Newbold L, Oliver A, Smyth T, et al. The seasonal structure of microbial communities in the Western English Channel. *Environ Microbiol*. 2009;11:3132–3139.

43. Teeling H, Fuchs BM, Bennke CM, Krüger K, Chafee M, Kappelmann L, et al. Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife*. 2016;5:e11888.

44. 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*. 2012;336:608–611.

45. Kirchman DL, Cottrell MT, Lovejoy C. The structure of bacterial communities in the western Arctic Ocean as revealed by pyrosequencing of 16S rRNA genes. *Environ Microbiol*. 2010;12:1132–1143.

46. Campbell BJ, Yu L, Heidelberg JF, Kirchman DL. Activity of abundant and rare bacteria in a coastal ocean. *Proc Natl Acad Sci*. 2011;108:12776–12781.

47. Fuhrman JA, Steele JA. Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquat Microb Ecol*. 2008;53:69–81.

48. Lewis KM, Dijken GL van, Arrigo KR. Changes in phytoplankton concentration now drive increased Arctic Ocean primary production. *Science*. 2020;369:198–202.

49. Arrigo KR, van Dijken GL. Continued increases in Arctic Ocean primary production. *Prog Oceanogr*. 2015;136:60–70.

50. Oziel L, Neukermans G, Ardyna M, Lancelot C, Tison J-L, Wassmann P, et al. Role for Atlantic inflows and sea ice loss on shifting phytoplankton blooms in the Barents Sea. *J Geophys Res Oceans*. 2017;122:5121–5139.

51. Nöthig E-M, Bracher A, Engel A, Metfies K, Niehoff B, Peeken I, et al. Summertime plankton ecology in Fram Strait—a compilation of long- and short-term observations. *Polar Res*. 2015;34:23349.

52. von Appen W-J, Waite AM, Bergmann M, Bienhold C, Boebel O, Bracher A, et al. Sea-ice derived meltwater stratification slows the biological carbon pump: results from continuous observations. *Nat Commun*. 2021;12:7309.

53. Carter-Gates M, Balestreri C, Thorpe SE, Cottier F, Baylay A, Bibby TS, et al. Implications of increasing Atlantic influence for Arctic microbial community structure. *Sci Rep*. 2020;10:19262.

54. Agogué H, Lamy D, Neal PR, Sogin ML, Herndl GJ. Water mass-specificity of bacterial communities in the North Atlantic revealed by massively parallel sequencing. *Mol Ecol*. 2011;20:258–274.

55. Woodgate RA. Increases in the Pacific inflow to the Arctic from 1990 to 2015, and insights into seasonal trends and driving mechanisms from year-round Bering Strait mooring data. *Prog Oceanogr*. 2018;160:124–154.

56. Bélanger S, Xie H, Krotkov N, Larouche P, Vincent WF, Babin M. Photomineralization of terrigenous dissolved organic matter in Arctic coastal waters from 1979 to 2003: Interannual variability and implications of climate change. *Global Biogeochem Cycles*. 2006;20.

57. Lorenson TD, Greinert J, Coffin RB. Dissolved methane in the Beaufort Sea and the Arctic Ocean, 1992–2009; sources and atmospheric flux. *Limnol Oceanogr*. 2016;61:S300–S323.

58. Shakhova N, Semiletov I, Salyuk A, Yusupov V, Kosmach D, Gustafsson Ö. Extensive methane venting to the atmosphere from sediments of the East Siberian Arctic Shelf. *Science*. 2010;327:1246–1250.

59. Bowman JS, Rasmussen S, Blom N, Deming JW, Rysgaard S, Sicheritz-Ponten T. Microbial community structure of Arctic multiyear sea ice and surface seawater by 454 sequencing of the 16S RNA gene. *ISME J*. 2012;6:11–20.

60. Boetius A, Anesio AM, Deming JW, Mikucki JA, Rapp JZ. Microbial ecology of the cryosphere: sea ice and glacial habitats. *Nat Rev Microbiol*. 2015;13:677–690.

61. Yergeau E, Michel C, Tremblay J, Niemi A, King TL, Wyglinski J, et al. Metagenomic survey of the taxonomic and functional microbial communities of seawater and sea ice from the Canadian Arctic. *Sci Rep*. 2017;7:42242.

62. Lofthus S, Bakke I, Greer CW, Brakstad OG. Biodegradation of weathered crude oil by microbial communities in solid and melted sea ice. *Marine Poll Bull*. 2021;172:112823.

63. Lantuit H, Overduin PP, Couture N, Wetterich S, Aré F, Atkinson D, et al. The Arctic coastal dynamics database: a new classification scheme and statistics on Arctic permafrost coastlines. *Estuaries and Coasts*. 2012;35:383–400.

64. Jones BM, Arp CD, Jorgenson MT, Hinkel KM, Schmutz JA, Flint PL. Increase in the rate and uniformity of coastline erosion in Arctic Alaska. *Geophys Res Lett*. 2009;36.

65. Vonk JE, Sánchez-García L, van Dongen BE, Alling V, Kosmach D, Charkin A, et al. Activation of old carbon by erosion of coastal and subsea permafrost in Arctic Siberia. *Nature*. 2012;489:137–140.

66. Wegner C, Bennett KE, Vernal A de, Forwick M, Fritz M, Heikkilä M, et al. Variability in transport of terrigenous material on the shelves and the deep Arctic Ocean during the Holocene. *Polar Res*. 2015;34.

67. Jong D, Bröder L, Tanski G, Fritz M, Lantuit H, Tesi T, et al. Nearshore zone dynamics determine pathway of organic carbon from eroding permafrost coasts. *Geophys Res Lett*. 2020;47:e2020GL088561.

68. Colatriano D, Tran PQ, Guéguen C, Williams WJ, Lovejoy C, Walsh DA. Genomic evidence for the degradation of terrestrial organic matter by pelagic Arctic Ocean Chloroflexi bacteria. *Commun Biol*. 2018;1:90.

69. Brenner D, Staley J, Krieg N. Classification of prokaryotic organisms and the concept of bacterial speciation. In: Bergey's Manual of Systematics of Archaea and Bacteria, eds (M. E. Trujillo, S. Dedysh, P. DeVos, B. Hedlund, P. Kämpfer, F. A. Rainey and W. B. Whitman). 2015. Springer-Verlag, New York, N.Y., pp 27–31.

70. Quast C, Pruesse E, Yilmaz P, Gerken J, Schweer T, Yarza P, et al. The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Res*. 2013;41:D590–D596.

71. Ludwig W, Strunk O, Westram R, Richter L, Meier H, Yadhukumar, et al. ARB: a software environment for sequence data. *Nucleic Acids Res*. 2004;32:1363–1371.

72. Yilmaz P, Parfrey LW, Yarza P, Gerken J, Pruesse E, Quast C, et al. The SILVA and "All-species Living Tree Project (LTP)" taxonomic frameworks. *Nucleic Acids Res*. 2014;42:643–648.

73. Klappenbach JA, Saxman PR, Cole JR, Schmidt TM. rrndb: the ribosomal RNA operon copy number database. *Nucleic Acids Res*. 2001;29:181–184.

74. Zaneveld JR, Lozupone C, Gordon JI, Knight R. Ribosomal RNA diversity predicts genome diversity in gut bacteria and their relatives. *Nucleic Acids Res*. 2010;38:3869–3879.

75. Lan Y, Rosen G, Hershberg R. Marker genes that are less conserved in their sequences are useful for predicting genome-wide similarity levels between closely related prokaryotic strains. *Microbiome*. 2016;4:18.

76. Case RJ, Boucher Y, Dahllöf I, Holmström C, Doolittle WF, Kjelleberg S. Use of 16S rRNA and rpoB genes as molecular markers for microbial ecology studies. *Appl Environ Microbiol*. 2007;73:278–288.

77. 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;1:1–6.

78. Olm MR, Crits-Christoph A, Diamond S, Lavy A, Matheus Carnevali PB, Banfield JF. Consistent metagenome-derived metrics verify and delineate bacterial species boundaries. *mSystems*. 2020;5:e00731-19.

79. von Meijenfeldt FAB, Arkhipova K, Cambuy DD, Coutinho FH, Dutilh BE. Robust taxonomic classification of uncharted microbial sequences and bins with CAT and BAT. *Genome Biology*. 2019;20:217.

80. Mirdita M, Steinegger M, Breitwieser F, Söding J, Levy Karin E. Fast and sensitive taxonomic assignment to metagenomic contigs. *J Bioinform*. 2021;37:3029–3031.

81. Hyatt D, Chen G-L, LoCascio PF, Land ML, Larimer FW, Hauser LJ. Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinform*. 2010;11:119.

82. Rho M, Tang H, Ye Y. FragGeneScan: predicting genes in short and error-prone reads. *Nucleic Acids Res*. 2010;38:e191.

83. 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;50:785–794.

84. Gilbert JA, Steele JA, Caporaso JG, Steinbrück L, Reeder J, Temperton B, et al. Defining seasonal marine microbial community dynamics. *ISME J*. 2012;6:298–308.

85. Portik DM, Brown CT, Pierce-Ward NT. Evaluation of taxonomic profiling methods for long-read shotgun metagenomic sequencing datasets. 2022. BioXriv. , 2022.01.31.478527

86. Huson DH, Albrecht B, Bağcı C, Bessarab I, Górska A, Jolic D, et al. MEGAN-LR: new algorithms allow accurate binning and easy interactive exploration of metagenomic long reads and contigs. *Biol Direct*. 2018;13:6.

87. Garber AI, Armbruster CR, Lee SE, Cooper VS, Bomberger JM, McAllister SM. SprayNPray: user-friendly taxonomic profiling of genome and metagenome contigs. *BMC Genom*. 2022;23:202.

88. Grinnell J. The niche-relationships of the California Thrasher. *The Auk*. 1917;34:427–433.

89. Elton CS. Animal ecology. 1927. Macmillan Co., New York.

90. Hutchinson GE. Concluding remarks. *Cold Spring Harbour symposium on quantitative biology*. 1957;22:415–427.

91. Hutchinson GE. An introduction to population biology. 1978. Yale Univ Press, New Haven, CT.

92. Ghiglione J-F, Galand PE, Pommier T, Pedrós-Alió C, Maas EW, Bakker K, et al. Pole-to-pole biogeography of surface and deep marine bacterial communities. *Proc Natl Acad Sci*. 2012;109:17633–17638.

93. Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EMS, Chisholm SW. Niche partitioning among prochlorococcus ecotypes along ocean-scale environmental gradients. *Science*. 2006;311:1737–1740.

94. Wang Z, Juarez DL, Pan J-F, Blinebry SK, Gronniger J, Clark JS, et al. Microbial communities across nearshore to offshore coastal transects are primarily shaped by distance and temperature. *Environ Microbiol*. 2019;21:3862–3872.

95. Herlemann DPR, Labrenz M, Jürgens K, Bertilsson S, Waniek JJ, Andersson AF. Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J*. 2011;5:1571–1579.

96. Alonso C, Gómez-Pereira P, Ramette A, Ortega L, Fuchs BM, Amann R. Multilevel analysis of the bacterial diversity along the environmental gradient Río de la Plata–South Atlantic Ocean. *Aquat Microb Ecol*. 2010;61:57–72.

97. Stanley SO, Morita RY. Salinity effect on the maximal growth temperature of some bacteria isolated from marine environments. *J Bacteriol*. 1968;95:169–173.

98. Schiraldi C, De Rosa M. Mesophilic Organisms. In: Encyclopedia of membranes, eds (Drioli, E., Giorno, L.). 2015. Springer, Berlin, Heidelberg, pp 1–2.

99. Reimers CE, Suess E. The partitioning of organic carbon fluxes and sedimentary organic matter decomposition rates in the ocean. *Mar Chem*. 1983;13:141–168.

100. Wright JJ, Konwar KM, Hallam SJ. Microbial ecology of expanding oxygen minimum zones. *Nat Rev Microbiol*. 2012;10:381–394.

101. Smriga S, Ciccarese D, Babbin AR. Denitrifying bacteria respond to and shape microscale gradients within particulate matrices. *Commun Biol*. 2021;4:1–9.

102. Garcia-Robledo E, Padilla CC, Aldunate M, Stewart FJ, Ulloa O, Paulmier A, et al. Cryptic oxygen cycling in anoxic marine zones. *Proc Natl Acad Sci*. 2017;114:8319–8324.

103. Márquez-Artavia A, Sánchez-Velasco L, Barton ED, Paulmier A, Santamaría-Del-Ángel E, Beier E. A suboxic chlorophyll-a maximum persists within the Pacific oxygen minimum zone off Mexico. *Deep Sea Res Part II: Top Stud Oceanogr*. 2019;169–170:104686.

104. Thomsen S, Kanzow T, Colas F, Echevin V, Krahmann G, Engel A. Do submesoscale frontal processes ventilate the oxygen minimum zone off Peru? *Geophys Res Lett*. 2016;43:8133–8142.

105. Tsementzi D, Wu J, Deutsch S, Nath S, Rodriguez-R LM, Burns AS, et al. SAR11 bacteria linked to ocean anoxia and nitrogen loss. *Nature*. 2016;536:179–183.

106. Berg JS, Ahmerkamp S, Pjevac P, Hausmann B, Milucka J, Kuypers MMM. How low can they go? Aerobic respiration by microorganisms under apparent anoxia. *FEMS Microbiol Rev*. 2022;46:fuac006.

107. Marchant HK, Ahmerkamp S, Lavik G, Tegetmeyer HE, Graf J, Klatt JM, et al. Denitrifying community in coastal sediments performs aerobic and anaerobic respiration simultaneously. *ISME J*. 2017;11:1799–1812.

108. Stramma L, Prince ED, Schmidtko S, Luo J, Hoolihan JP, Visbeck M, et al. Expansion of oxygen minimum zones may reduce available habitat for tropical pelagic fishes. *Nature Clim Change*. 2012;2:33–37.

109. Robinson C. Microbial respiration, the Engine of ocean deoxygenation. *Front Mar Sci*. 2019;5.

110. Breitburg D, Levin LA, Oschlies A, Grégoire M, Chavez FP, Conley DJ, et al. Declining oxygen in the global ocean and coastal waters. *Science*. 2018;359:eaam7240.

111. Finkel OM, Béjà O, Belkin S. Global abundance of microbial rhodopsins. *ISME J*. 2013;7:448–451.

112. Gómez-Consarnau L, Akram N, Lindell K, Pedersen A, Neutze R, Milton DL, et al. Proteorhodopsin phototrophy promotes survival of marine bacteria during starvation. *PLoS Biol*. 2010;8:e1000358.

113. Sabehi G, Kirkup BC, Rozenberg M, Stambler N, Polz MF, Béjà O. Adaptation and spectral tuning in divergent marine proteorhodopsins from the eastern Mediterranean and the Sargasso Seas. *ISME J*. 2007;1:48–55.

114. Morris JJ, Lenski RE, Zinser ER. The Black Queen Hypothesis: evolution of dependencies through adaptive gene loss. *mBio*. 2012;3:e00036-12.

115. Avcı B, Krüger K, Fuchs BM, Teeling H, Amann RI. Polysaccharide niche partitioning of distinct Polaribacter clades during North Sea spring algal blooms. *ISME J*. 2020;14:1369–1383.

116. Krüger K, Chafee M, Ben Francis T, Glavina del Rio T, Becher D, Schweder T, et al. In marine Bacteroidetes the bulk of glycan degradation during algae blooms is mediated by few clades using a restricted set of genes. *ISME J*. 2019;13:2800–2816.

117. Chen J, Robb CS, Unfried F, Kappelmann L, Markert S, Song T, et al. Alpha- and beta-mannan utilization by marine Bacteroidetes. *Environm Microbiol*. 2018;20:4127–4140.

118. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. *Nature*. 2021;596:583–589.

119. Gómez-Pereira PR, Fuchs BM, Alonso C, Oliver MJ, van Beusekom JEE, Amann R. Distinct flavobacterial communities in contrasting water masses of the North Atlantic Ocean. *ISME J*. 2010;4:472–487.

120. Milici M, Vital M, Tomasch J, Badewien TH, Giebel H-A, Plumeier I, et al. Diversity and community composition of particle-associated and free-living bacteria in mesopelagic and bathypelagic Southern Ocean water masses: Evidence of dispersal limitation in the Bransfield Strait. *Limnol and Oceanogr*. 2017;62:1080–1095.

121. Jain C, Rodriguez-R LM, Phillippy AM, Konstantinidis KT, Aluru S. High throughput ANI analysis of 90K prokaryotic genomes reveals clear species boundaries. *Nat Commun*. 2018;9:5114.

122. Rodriguez-R LM, Jain C, Conrad RE, Aluru S, Konstantinidis KT. Reply to: “Re-evaluating the evidence for a universal genetic boundary among microbial species”. *Nat Commun*. 2021;12:4060.

123. Needham DM, Sachdeva R, Fuhrman JA. Ecological dynamics and co-occurrence among marine phytoplankton, bacteria and myoviruses shows microdiversity matters. *ISME J*. 2017;11:1614–1629.

124. Chafee M, Fernàndez-Guerra A, Buttigieg PL, Gerdts G, Eren AM, Teeling H, et al. Recurrent patterns of microdiversity in a temperate coastal marine environment. *ISME J*. 2018;12:237–252.

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

126. Cho J-C, Giovannoni SJ. Cultivation and growth characteristics of a diverse group of oligotrophic marine Gammaproteobacteria. *Appl Environ Microbiol*. 2004;70:432–440.

127. Takai K, Nunoura T, Ishibashi J, Lupton J, Suzuki R, Hamasaki H, et al. Variability in the microbial communities and hydrothermal fluid chemistry at the newly discovered Mariner hydrothermal field, southern Lau Basin. *J Geophys Res Biogeo*. 2008;113:G02031.

128. Nakagawa S, Takai K, Inagaki F, Chiba H, Ishibashi J, Kataoka S, et al. Variability in microbial community and venting chemistry in a sediment-hosted backarc hydrothermal system: Impacts of subseafloor phase-separation. *FEMS Microbiol Ecol*. 2005;54:141–155.

129. Fuhrman JA, Hewson I, Schwalbach MS, Steele JA, Brown MV, Naeem S. Annually reoccurring bacterial communities are predictable from ocean conditions. *Proc Natl Acad Sci*. 2006;103:13104–13109.

130. Hagen R, Plüß A, Ihde R, Freund J, Dreier N, Nehlsen E, et al. An itegrated marine data collection for the German Bight – part II: tides, salinity and waves (1996–2015 CE). *Earth Syst Sci Data*. 2021;13:2573–2594.

131. Pedrós-Alió C. Marine microbial diversity: can it be determined? *Trends Microbiol*. 2006;14:257–263.

132. Lennon JT, Jones SE. Microbial seed banks: the ecological and evolutionary implications of dormancy. *Nat Rev Microbiol*. 2011;9:119–130.

133. Baas Becking LGM. Geobiologie of inleiding tot de milieukunde. 1934. W.P. Van Stockum & Zoon, the Hague, Netherlands.

134. Hutchinson GE. The paradox of the Plankton. *Am Nat*. 1961;95:137–145.

135. Kashtan N, Roggensack SE, Rodrigue S, Thompson JW, Biller SJ, Coe A, et al. Single-cell genomics reveals hundreds of coexisting subpopulations in wild Prochlorococcus. *Science*. 2014;344:416–420.

136. Rodriguez-R LM, Conrad RE, Feistel DJ, Viver T, Rosselló-Móra R, Konstantinidis KT. A natural definition for a bacterial strain and clonal complex. 2022. BioXriv. , 2022.06.27.497766

137. Karsenti E, Acinas SG, Bork P, Bowler C, Vargas CD, Raes J, et al. A holistic approach to marine eco-systems biology. *PLOS Biology*. 2011;9:e1001177.

138. Duarte CM. Seafaring in the 21st century: the Malaspina 2010 circumnavigation expedition. *Limnol Oceanogr Bull*. 2015;24:11–14.

Appendix

Additional co-author and first author publications

environmental microbiology

Environmental Microbiology (2021) **23**(1), 448–463 doi:10.1111/1462-2920.15331

Diversity and biomass dynamics of unicellular marine fungi during a spring phytoplankton bloom

Taylor Priest,[1] **Bernhard Fuchs**,[1] **Rudolf Amann**[1] **and Marlis Reich**[2*]

[1]*Department of Molecular Ecology, Max Planck Institute for Marine Microbiology, Bremen, Germany.*

[2]*Molecular Ecology Group, FB2, University of Bremen, Bremen, Germany.*

Summary

Microbial communities have important functions during spring phytoplankton blooms, regulating bloom dynamics and processing organic matter. Despite extensive research into such processes, an in-depth assessment of the fungal component is missing, especially for the smaller size fractions. We investigated the dynamics of unicellular mycoplankton during a spring phytoplankton bloom in the North Sea by 18S rRNA gene tag sequencing and a modified CARD-FISH protocol. Visualization and enumeration of dominant taxa revealed unique cell count patterns that varied considerably over short time scales. The Rozellomycota *sensu lato (s.l.)* reached a maximum of 10^5 cells L^{-1}, being comparable to freshwater counts. The abundance of Dikarya surpassed previous values by two orders of magnitude (10^5 cells L^{-1}) and the corresponding biomass (maximum of 8.9 mg C m^{-3}) was comparable to one reported for filamentous fungi with assigned ecological importance. Our results show that unicellular fungi are an abundant and, based on high cellular ribosome content and fast dynamics, active part of coastal microbial communities. The known ecology of the visualized taxa and the observed dynamics suggest the existence of different ecological niches that link primary and secondary food chains, highlighting the importance of unicellular fungi in food web structures and carbon transfer.

Introduction

The presence of fungi in marine ecosystems has long been known, but their significance largely overlooked. Only recently, through environmental sequencing surveys, has their ubiquitous distribution and high diversity been revealed (Wang *et al*., 2014; Zhang *et al*., 2015; Comeau *et al*., 2016; Taylor and Cunliffe, 2016; Duan *et al*., 2018). Such studies have provided a valuable base-line understanding of marine fungi but much more work is required before we can understand their ecological roles and contribution to biogeochemical processes. In particular, information regarding cellular abundances, biomass and dynamics of dominant taxa over short-time scales is lacking (Grossart *et al*., 2019).

Pelagic fungi (mycoplankton) in the coastal environment have been shown to represent a diverse range of taxonomic groups and are able to accumulate considerable biomass, comparable to prokaryotes (Gutiérrez *et al*., 2010). The structure of these communities appears to vary with latitude and local influences, with evidence suggesting a dominance by members of the Dikarya (Ascomycota and Basidiomycota) or the Chytridiomycota (Comeau *et al*., 2016; Hassett *et al*., 2016; Tisthammer *et al*., 2016; Duan *et al*., 2018). Time-series data have also highlighted the impact of seasons on community composition and more specifically, changes in nutrients, temperature and, with respect to certain taxa, phytoplankton (Taylor and Cunliffe, 2016; Duan *et al*., 2018; Banos *et al*., 2020). To understand these driving factors further, a few studies have tried to elucidate ecological roles for the dominant fungal taxa and revealed the presence of parasitism and saprotrophism within the water column (Gutiérrez *et al*., 2016; Scholz *et al*., 2016; Cunliffe *et al*., 2017). Cunliffe *et al*. (2017) isolated several annually recurring mycoplankton strains from the English Channel and showed their ability to degrade phytoplankton-derived polysaccharides. In the coastal Arctic environment, Hassett and Gradinger (2016) used epifluorescence microscopy to visualize fungal parasitism of diatoms in sea-ice and seawater while Ishida *et al* (2015) described a great fungal taxonomic diversity associated with single freshwater phytoplankton cells. These insights would suggest that mycoplankton are active members of coastal ecosystems and likely perform important ecological functions during phytoplankton blooms. However, until now, none of the studies provide

Received 11 June, 2020; revised 11 November, 2020; accepted 12 November, 2020. *For correspondence. Tel. +49 (0)421 218 62825; Fax: +49 (0)421 218 98 62825; E-mail reich@uni-bremen.de.

Not digested: algal glycans move carbon dioxide into the deep-sea

Silvia Vidal-Melgosa[a,b,1,*], Matija Lagator[a,1,2], Andreas Sichert[a,b,3], Taylor Priest[a], Jürgen Pätzold[b], Jan-Hendrik Hehemann[a,b,*]

[a] Max Planck Institute for Marine Microbiology, 28359 Bremen, Germany.

[b] MARUM - Center for Marine Environmental Sciences, University of Bremen, 28359 Bremen, Germany.

[1] S.V.-M and M.L. contributed equally to this work.

[2] Present address: Department of Chemistry, Manchester Institute of Biotechnology, University of Manchester, Manchester, M1 7DN, UK.

[3] Present address: Institute of Molecular Systems Biology, ETH Zurich, 8093 Zurich, Switzerland.

*Corresponding authors. Email address: svidal@mpi-bremen.de; jhhehemann@marum.de.

Abstract

Marine algae annually synthesize gigatons of glycans from carbon dioxide, exporting it within sinking particles into the deep-sea and underlying sea floor, unless those glycans are digested before by bacteria. Identifying algal glycans in the ocean remains challenging with the molecular resolution of conventional analytic techniques. Whether algal glycans are digested by heterotrophic bacteria during downward transport, before they can transfer carbon dioxide from the ocean surface into the deep-sea or the sea floor, remains unknown. In the Red Sea Shaban Deep, where at 1500 m water depth a brine basin acts as a natural sediment trap, we found its high salt and low oxygen concentration accumulated and preserved exported algal glycans for the past 2500 years. By using monoclonal antibodies specific for glycan structures, we detected fucose-containing sulfated polysaccharide, β-glucan, β-mannan and arabinogalactan glycans, synthesized by diatoms, coccolithophores, dinoflagellates and other algae living in the sunlit ocean. Their presence in deep-sea sediment demonstrates these algal glycans were not digested by bacteria. Instead they moved carbon dioxide from the surface ocean into the deep-sea, where it will be locked away from the atmosphere at least for the next 1000 years. Considering their global synthesis, quantity and stability against degradation during transport through the water column, algal glycans are agents for carbon sequestration.

Keywords: Marine sediment; Biological carbon pump; Algae polysaccharide; Glycan; Carbon sequestration; Extracellular matrix.

Manuscript under review in PNAS Journal

High abundance of hydrocarbon-degrading *Alcanivorax* in plumes of hydrothermally active volcanoes in the South Pacific Ocean

Bledina Dede[1], Taylor Priest[1], Wolfgang Bach[2,3], Rudolf Amann[1], Anke Meyerdierks[1*]

[1] Max Planck Institute for Marine Microbiology, Bremen, Germany

[2] MARUM, Center for Marine Environmental Sciences, University of Bremen, Bremen, Germany

[3] Geoscience Department, University of Bremen, Bremen, Germany

***Corresponding author:**

Anke Meyerdierks

ameyerdi@mpi-bremen.de

ABSTRACT

Species within the genus *Alcanivorax* are well known hydrocarbon-degraders that propagate quickly in oil spills and natural oil seepage. They are also inhabitants of the deep-sea and have been found in several hydrothermal plumes. However, an in depth analysis of deep-sea *Alcanivorax* is currently lacking. In this study, we used multiple culture-independent techniques to analyze the microbial community composition of hydrothermal plumes in the Northern Tonga arc and Northeastern Lau Basin focusing on the autecology of *Alcanivorax*. The hydrothermal vents feeding the plumes are hosted in an arc volcano (Niua), a rear-arc caldera (Niuatahi) and the Northeast Lau Spreading Centre (Maka). Fluorescence *in situ* hybridization revealed that *Alcanivorax* dominated the community at two sites (1210 – 1565 mbsl), reaching up to 48% relative abundance and 3.5 x 10^4 cells/ml. Through 16S rRNA gene and metagenome analyses, we identified that this pattern was driven by two *Alcanivorax* species in the plumes of Niuatahi and Maka. Despite no indication for hydrocarbon presence in the plumes of these areas, a high expression of genes involved in hydrocarbon-degradation was observed. We hypothesize that the high abundance and gene expression of *Alcanivorax* is likely due to yet undiscovered hydrocarbon seepage from the seafloor, potentially resulting from recent volcanic activity in the area. Chain-length and complexity of hydrocarbons, and water depth could be driving niche partitioning in *Alcanivorax*.

Manuscript under review in ISME Journal

Acknowledgements

Firstly, I would like to thank Prof. Dr. A. Murat Eren, Dr. Silvia G. Acinas and PD. Dr. Bernhard M. Fuchs for agreeing to review this thesis and be a part of my examination board. An additional thanks to Jan-Hendrik Hehemann, Margot Bligh and Mahum Farhan for agreeing to be members of my examination board.

I would like to extend the most heartfelt and sincere thanks to Bernhard and Rudi. Not only have they supported, advised and guided me throughout the course of this thesis, but they gave me the opportunity and means to study what I am truly passionate about. For that, I will be forever grateful. A huge thanks to everybody in the FACS group and Molecular Ecology department for welcoming and supporting me and being the best the colleagues anyone could ask for. In particular, I would like to thank Mirja, Jörg and Kathrin for their support and guidance in the lab over the past few years. I also would like to extend the most sincere thanks to Coto, who has taught me so much about bioinformatics and who was always available to answer questions and discuss ideas with me. Thanks also to Jan Brüwer for offering to help with the German translation of the summary for this thesis.

I am also incredibly grateful to all of the collaborators and colleagues that contributed to and supported me in the research presented in this thesis. In particular, Anni Heins, Bruno Hüttel, Silvia Vidal-Melgosa, Jens Harder and Matthias Wietz. A special, extended thanks to Matthias for providing me with the opportunity to undertake the research in Chapter IV and Silvia, for teaching and training me on carbohydrate analysis that made Chapter V possible.

An important thanks to everybody behind the IMPRS MarMic program. The program has truly been life-changing and I feel incredibly grateful and blessed to have been a part of it.

I have also been so lucky to have the unwavering support of my family back in England. I am thankful every day to have them in my life. In addition, I am humbled and so grateful to have such a supportive friendship group in Bremen that have made the past few years special.

Finally, it goes far beyond what I can express in words, but this thesis would not have been possible without the love, care and support from my wife, Bledina. You make me smile, laugh and feel loved everyday and you have helped me to grow and develop so much as a person over the past few years. I am eternally grateful to have you in my life. Although I doubt he will read this thesis, I must also express my gratitude to our beautiful dog Nox, who brings an immeasurable amount of happiness into my life.

Universität Bremen

Ort, Datum: _____

Versicherung an Eides Statt

Ich, Taylor Priest

.

versichere an Eides Statt durch meine Unterschrift, dass ich die vorstehende Arbeit selbständig und ohne fremde Hilfe angefertigt und alle Stellen, die ich wörtlich dem Sinne nach aus Veröffentlichungen entnommen habe, als solche kenntlich gemacht habe, mich auch keiner anderen als der angegebenen Literatur oder sonstiger Hilfsmittel bedient habe.
Ich versichere an Eides Statt, dass ich die vorgenannten Angaben nach bestem Wissen und Gewissen gemacht habe und dass die Angaben der Wahrheit entsprechen und ich nichts verschwiegen habe.
Die Strafbarkeit einer falschen eidesstattlichen Versicherung ist mir bekannt, namentlich die Strafandrohung gemäß § 156 StGB bis zu drei Jahren Freiheitsstrafe oder Geldstrafe bei vorsätzlicher Begehung der Tat bzw. gemäß § 161 Abs. 1 StGB bis zu einem Jahr Freiheitsstrafe oder Geldstrafe bei fahrlässiger Begehung.

_____
Ort, Datum Unterschrift

Changes made after the review process

1) A hyphen was introduced into the title, from “high latitude” to “high-latitude”