Polysaccharide utilization loci and associated genes in marine *Bacteroidetes* - compositional diversity and ecological relevance

Dissertation
zur Erlangung des Doktorgrades
der Naturwissenschaften
- Dr. rer. nat. -

dem Fachbereich 2 Biologie/Chemie
der Universität Bremen
vorgelegt von

Karen Krüger

Bremen, Januar 2019

Polysaccharide utilization loci and associated genes in marine *Bacteroidetes* - compositional diversity and ecological relevance

A thesis submitted to University of Bremen in
partial fulfillment for the degree of
DOCTOR OF SCIENCE (Doktor der Naturwissenschaften)

University of Bremen
Faculty 2 Biology/Chemistry

Karen Krüger

Bremen, January 2019

Die vorliegende Doktorarbeit wurde im Rahmen des Programms *International Max Planck Research School of Marine Microbiology* (MarMic) in der Zeit von Oktober 2014 bis Februar 2019 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 October 2014 to February 2019.

Gutachter : Prof. Dr. Rudolf Amann
Gutachter : Prof. Dr. Carol Arnosti

Prüfer : Prof. Dr. Michael Friedrich
Prüfer : Dr. Jan-Hendrik Hehemann

Tag des Promotionskolloquiums: 06.03.2019

4 3 2 1

Summary

The synthesis of marine organic carbon compounds by photosynthetic macroalgae, microalgae (phytoplankton) and bacteria provide a basis for life in the ocean. In marine surface waters this primary production is largely dominated by microalgae and is especially pronounced during spring phytoplankton blooms. During and after these often diatom-dominated blooms, increased amounts of organic matter are released into the surrounding waters. Here, the organic matter, rich in polysaccharides, can trigger blooms of heterotrophic bacteria. Marine members of the *Bacteroidetes* are consistently found related to such bloom events. These bacteria are regularly detected as the first responders to thrive after phytoplankton spring blooms in temperate coastal regions and are often equipped with a variety of polysaccharide utilization gene clusters. These gene clusters, termed polysaccharide utilization loci (PULs), encode enzymes for the extracellular hydrolysis of polysaccharides and the subsequent uptake of oligosaccharides into the periplasm, where they are shielded from competing bacteria. This mechanism allows for rapid uptake and substrate hoarding, and thus could be one reason why *Bacteroidetes* are often seen as the first responders of the bacterioplankton community.

The investigation of the so far largely unknown diversity and the ecological relevance of PULs in marine *Bacteroidetes* was the major goal of the work presented here. We could show that genomes of *Bacteroidetes* isolates from the North Sea, with free-living to micro- and macro-algae associated lifestyles, harboured a variety of these loci predicted to target in total 18 different substrate classes. Overall PUL repertoires of these isolates showed considerable intra-genus and inter-genus, variations suggesting that *Bacteroidetes* species harbour distinct glycan niches, independent of their phylogenetic relationships. By investigating the PUL repertoires of uncultured free-living *Bacteroidetes* during three consecutive years of spring phytoplankton blooms at the North Sea island of Helgoland, I could further reveal that the set of targeted substrates during these bloom events was dominated by only five of the substrate classes targeted by the isolates. These were the diatom storage polysaccharide laminarin, alpha-glucans, alginates, as well as substrates rich in alpha-mannans and sulfated xylans. In addition to this constrained set of substrate classes targeted by the free-living *Bacteroidetes* community, I could show

that the species diversity during these blooms was limited and dominated by only 27 abundant and recurrent species that carried a limited number of abundant PULs. The majority of these PULs were targeting laminarin and alpha-glucan substrates, which were likely targeted during the entire time of the blooms. The less frequent PULs, targeting alpha-mannans and sulfated xylans, were predominantly detected during mid- and late- bloom phases, suggesting a relevance of these two substrate classes in the later phases of phytoplankton blooms. Overall these findings highlight the recurrence of a few specialized *Bacteroidetes* species and the environmental relevance of specific polysaccharide substrate classes during spring phytoplankton blooms. However, for some of these substrate classes the origin, structural details and their abundance during blooms are as yet largely unknown. To further shed light on the polysaccharide niches of abundant key-players, these findings can serve as a guide for future laboratory studies.

Zusammenfassung

Das Leben im Ozean baut maßgeblich auf der Synthese von organischen Kohlenstoffverbindungen durch photosynthetische Makroalgen, Mikroalgen (Phytoplankton) sowie Bakterien auf. In marinen Oberflächengewässern wird diese Primärproduktion weitestgehend von Mikroalgen dominiert und zeigt besonders während der Phytoplanktonblüte im Frühjahr eine deutliche Ausprägung. Während und nach diesen oft von Diatomeen dominierten Blüten werden vermehrt organische Substanzen, die reich an Polysacchariden sind, in das umliegende Wasser freigesetzt. Diese organischen Substanzen lösen häufig sekundäre Blüten heterotropher Bakterien aus, welche oft von marinen Vertretern der *Bacteroidetes* dominiert werden. In gemäßigten Küstenregionen werden diese Bakterien regelmäßig als erste dominierende Gruppe nach der Phytoplankton-Blüte detektiert. Marine Vertreter der *Bacteroidetes* sind oft mit einer Vielzahl an Genclustern für den Polysaccharidabbau ausgestattet. Diese Gencluster, die als "polysaccharide utilization loci" (PULs) bezeichnet werden, kodieren für Enzyme, die die extrazelluläre Hydrolyse von Polysacchariden und die anschließende Aufnahme von Oligosacchariden in das Periplasma ermöglichen. Hier sind Oligosaccharide von konkurrierenden Bakterien abgeschirmt und können weiter hydrolisiert werden. Dieser Mechanismus ermöglicht eine schnelle Aufnahme von Substraten und könnte daher ein Grund dafür sein, warum *Bacteroidetes* häufig als erste Gruppe der Bakterioplanktongemeinschaft in sekundären Blüten auftreten.

Das Hauptziel der hier vorgestellten Arbeit war die Untersuchung der bislang weitgehend unbekannten Vielfalt und der ökologischen Relevanz von PULs in marinen *Bacteroidetes*. Es konnte gezeigt werden, dass Genome von *Bacteroidetes*-Isolaten aus der Nordsee, mit frei lebenden oder Mikro- und Makroalgen assoziierten Lebensstilen, eine Vielzahl dieser Gencluster beherbergten. Für diese wurde der Abbau von insgesamt 18 verschiedenen Polysaccharidsubstratklassen vorhergesagt. Die PUL-Repertoires dieser Isolate zeigten ferner eine beträchtliche Variation innerhalb sowie zwischen den *Bacteroidetes*-Gattungen. Dies deutet darauf hin, dass *Bacteroidetes*-Arten unabhängig von ihrer phylogenetischen Beziehung unterschiedliche Nischen im Polysaccharidabbau aufweisen. Zusätzlich zu den *Bacteroidetes*-Isolaten wurden PUL-Repertoires von unkultivierten,

frei-lebenden *Bacteroidetes* während dreier aufeinanderfolgender Frühjahrsphytoplanktonblüten vor der Nordseeinsel Helgoland untersucht. Es konnte festgestellt werden, dass das Set an Zielsubstraten dieser PULs von nur fünf Substratklassen dominiert wurde, die auch schon in den Isolaten nachgewiesen werden konnten. Bei diesen handelt es sich um Alpha-Glucane, Alginate, das Diatomeen-Speicherpolysaccharid Laminarin sowie Substrate, die reich an Alpha-Mannanen und sulfatierten Xylanen waren. Zusätzlich zu diesem beschränkten Set an Substratklassen, die von der frei-lebenden *Bacteroidetes*-Gemeinschaft abgebaut werden, konnte dargelegt werden, dass die Artenvielfalt während der Blüte begrenzt war und von nur wenigen, wiederkehrenden Arten dominiert wurde, welche eine eingeschränkte Anzahl von PULs trugen. Die Mehrheit dieser PULs zielte auf Laminarin und Alpha-Glucansubstrate ab, die wahrscheinlich während der gesamten Blüte abgebaut wurden. Die weniger häufigen PULs, die für den Abbau von Alpha-Mannanen und sulfatierten Xylanen kodieren, wurden überwiegend in mittleren und späten Blütephasen nachgewiesen. Dies lässt auf eine Relevanz dieser beiden Substrate in den späten Phasen der Phytoplanktonblüte schließen. Insgesamt belegen die Ergebnisse dieser Arbeit das wiederkehrende Vorkommen von nur wenigen *Bacteroidetes*-Arten und die hohe Umweltrelevanz bestimmter Polysaccharidsubstratklassen während Phytoplanktonblüten im Frühjahr. Der Ursprung, die Struktur und die Abundanz während der Blüte ist für einige dieser Substratklassen jedoch noch weitgehend unbekannt. Um die Polysaccharid-Nischen verschiedener Schlüsselorganismen weiter zu beleuchten, können diese Ergebnisse zukünftig als Leitfaden für weiterführende Laboruntersuchungen dienen.

Abbreviations

AA auxiliary activity

AAI average amino acid identity

ABC ATP-binding cassette

ANI average nucleotide identity

BPLR Boreal Polar region

CARD-FISH catalysed reporter deposition-FISH

CAZyme carbohydrate active enzyme

CBM carbohydrate-binding module

CE carbohydrate esterase

COGITO Coastal Microbe Genomic and Taxonomic Observatory

DAPI 4',6-diamidino-2-phenylindole

DNA deoxyribonucleic acid

DOC dissolved organic carbon

DOE-JGI Department of Energy Joint Genome Institute

DOM dissolved organic matter

FCSP fucose-containing sulphated polysaccharides

FDR false discovery rate

FISH fluorescence in situ hybridisation

FLA fluorescently labelled

GFBio German Federation for Biological Data

GH glycoside hydrolase

GT glycosyltransferase

GTDB genome taxonomy database

HCR-FISH hybridisation chain reaction-FISH

HMW high-molecular-weight

HPAEC-PAD high performance anion exchange chromatography with pulsed amperometric detection

HTS high-throughput sequencing

iBAQ intensity-based absolute quantification

LGT lateral gene transfer

LMW low-molecular-weight

MAG metagenome-assembled genome

MDA multiple-displacement amplification

MFS major facilitator superfamily

MIMAS Microbial Interactions in Marine Systems

MIxS Minimal Information about any (X) Sequence

MS mass spectrometry

NAST North Atlantic Subtropical

NPP net primary production

NSAF normalized spectral abundance factors

OLC overlap layout consensus

OM organic matter

ORF open reading frame

OTU operational taxonomic unit

PCR polymerase chain reaction

PL polysaccharide lyase

POM particulate organic matter

PUL polysaccharide utilization locus

PULs polysaccharide utilization loci

PUL-DB PUL database

riBAQ relative iBAQ

RPKM reads per kilobase million

rRNA ribosomal ribonucleic acid

SNVs Single nucleotide variants

Sus starch utilization system

TBDR TonB-dependent receptor

TBDT TonB-dependent transporter

TEP transparent exopolymer particles

TRAP tripartite ATP-independent periplasmic

Contents

Chapter 1

Introduction

1.1 The ocean carbon cycle

While the element carbon makes up only about 0.17% by weight of Earth (Allègre et al., 2001), it constitutes the major building block of life. In the oceans, macroalgae, phytoplankton and photosynthetic bacteria use light energy to convert simple inorganic carbon dioxide into biomass consisting of manifold organic carbon compounds of varying complexity. This oceanic synthesis of carbon compounds, also termed oceanic primary production, contributes about half of the total net primary production (NPP) on Earth (Field et al., 1998). Additional carbon fixation in the ocean comes from light independent chemosynthesis by deep-sea bacteria, and from bacterial conversion of compounds such as carbon monoxide or methane into biomass. All of these carbon fixation processes provide a basis for oceanic life and fuel several larger carbon pools. Fluxes between these major carbon pools, such as the recycling and remineralization of dissolved organic matter (DOM) by heterotrophic bacteria, are important factors in marine carbon cycling (Fig. 1.1).

After fossil sedimentary organic carbon and dissolved inorganic carbon, DOM is the third largest carbon pool in the marine environment. It is estimated at about 662 Pg C, which is about 200 times the carbon present in marine biomass (Hansell et al., 2009). About 47 Pg C of this are attributed to the epipelagic zone (0-200 m depth; Hansell et al. (2009)). DOM is released by phytoplankton and other organisms as excretion products, and during decay or viral lysis (Fig. 1.1). About half of the DOM is recycled in the microbial loop by heterotrophic bacteria (Azam, 1998). These bacteria are capable of using the organic matter (OM) released by phytoplankton and respire it to CO_2. The low-molecular-weight (LMW) fraction of the DOM can be passively transported into the

bacterial cell through transmembrane porins or actively via transporters, while high-molecular-weight (HMW) DOM larger than 600 Da requires special uptake mechanisms and enzymatic hydrolysis before uptake into the cell (Weiss et al., 1991). The recycling and remodelling of DOM by bacteria is considered to be fuelling a pool of recalcitrant DOM, which contains organic molecules resistant to further microbial degradation. This was postulated in the concept of the microbial carbon pump (Jiao et al., 2010), which considers the storage and accumulation of this recalcitrant DOM in the ocean (Fig. 1.1). While recalcitrant DOM exists in the ocean, it has been shown that actually only a small fraction of DOM produced by microbes is recalcitrant DOM (Osterholz et al., 2015), whereas most is readily accessible as polysaccharides or proteins. Thus, the accumulation of recalcitrant DOM is more likely the result of low concentrations of molecules within this pool, which makes these molecules inaccessible for microbes (Arrieta et al., 2015).

Apart from DOM, particulate organic matter (POM) is released by phytoplankton during zooplankton grazing or cell lysis. The delineation of DOM and POM is an operational rather than biological one, and is based on size separation. While DOM is defined as the size range passing a filter of usually 0.2-0.7 μm, POM is defined as everything retained on these filters (Dittmar & Stubbins, 2014). POM consists of both living and non-living OM and a fraction of it, when partially degraded, can become part of the DOM pool, while DOM can form POM through aggregation. In contrast to the microbial carbon pump, the biological carbon pump (Turner, 2015) contributes to carbon sequestration by vertical export of POM to the bottom of the ocean by sinking processes. Overall though, DOM is the more abundant organic carbon reservoir in the marine realm and plays an important role in sustaining the heterotrophic bacterioplankton community in the ocean.

1.2 Phytoplankton blooms in temperate coastal regions

The cycling of carbon and the interplay of phytoplankton and heterotrophic bacteria is particularly pronounced during times of high productivity, such as during seasonal phytoplankton blooms. These annually recurring phenomena in temperate coastal regions are triggered by rising temperatures, increasing sunlight intensities and the availability of replenished nutrients. All these factors allow high primary productivity and thus boost phytoplankton growth, sometimes to such high cell densities in such large areas that these blooms can be observed from space (Fig. 1.2). Spring phytoplankton blooms in temperate coastal regions are usually dominated by diatoms, which require silicate for the formation of their outer cell wall (frustule). Diatoms are estimated to be responsible for about 40-45% of total oceanic NPP (Armbrust, 2009) and with that make up about

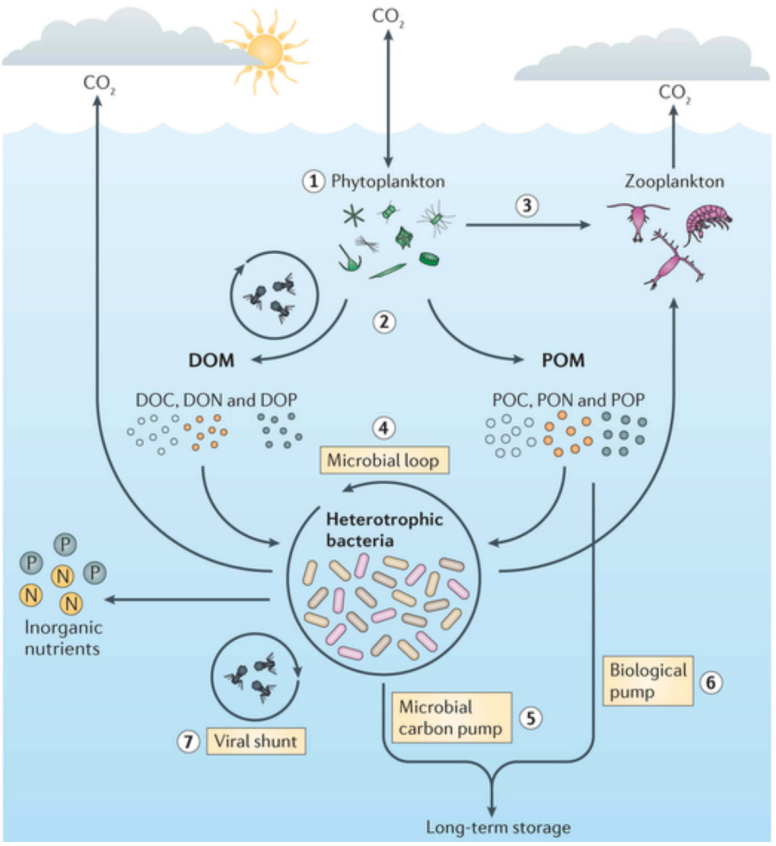


Figure 1.1: Bacterial transformation of phytoplankton-derived organic matter. (1) Conversion of inorganic carbon to organic carbon by photosynthetic phytoplankton species. (2) The release of dissolved organic matter (DOM) and particulate organic matter (POM) from phytoplankton. (3) Consumption of phytoplankton biomass by zooplankton grazers. (4) The mineralization and recycling of organic matter by diverse heterotrophic bacteria (known as the microbial loop). (5) The microbial carbon pump: transformation of organic carbon into recalcitrant dissolved organic carbon (DOC) that resists further degradation and is sequestered in the ocean for thousands of years. (6) The biological carbon pump: export of phytoplankton-derived POM from the surface oceans to deeper depths via sinking. (7) The viral shunt: contributions of viral-mediated cell lysis to the release of dissolved and particulate matter from both the phytoplankton and bacterial pools. (Adapted from Buchan et al. (2014) with permission by Springer Nature Customer Service Centre GmbH).

one-fifth of global primary production. Thus they contribute the largest single source of OM input to the marine system. Phytoplankton blooms can last for a few days up to several weeks and often display a succession of different species of diatoms or other phytoplankton species during this time. The subsequent decline of blooms can either be caused by nutrient limitation or competition for nutrients ("bottom up" control), or grazing and viral lysis ("top down" control), or it is a combination of both factors.

During phytoplankton blooms and during their decline, high amounts of fresh, readily accessible OM are released into the waters. This can then trigger secondary blooms of heterotrophic bacterioplankton (Landa et al., 2016, Needham & Fuhrman, 2016, Taylor et al., 2014, Teeling et al., 2012, Yang et al., 2015). Despite the high intrinsic dynamics

of phytoplankton bloom events, such as variations in phytoplankton species composition and variability in environmental parameters, a limited number of heterotrophic bacterial clades is consistently associated with these phytoplankton bloom events (Buchan et al., 2014, Teeling et al., 2012).

Figure 1.2: Satellite image of a phytoplankton bloom in the German Bight, North Sea. The yellow circle indicates the location of Helgoland. ©NASA images courtesy Jeff Schmaltz, MODIS Rapid Response Team, Goddard Space Flight Center.

1.3 Secondary blooms of heterotrophic bacterioplankton

Long-term ecological research stations and the often high-resolution time-series sampling conducted at these stations are an important step towards understanding microbial population patterns and their temporal as well as spatial distributions. At the San Pedro Ocean Time-series station, Fuhrman and colleagues have detected very consistent seasonal patterns of the microbial community (Chow et al., 2013, Cram et al., 2014, Fuhrman et al., 2015). Samples taken during the same month each year showed very similar community composition to each other over a time frame of 10 years, indicating that community composition follows seasonal patterns with some interannual variability (Fuhrman et al., 2015). This indicates a long-term stable but seasonal behaviour of the microbial community. While this study focused on overall seasonal patterns, samples from phytoplankton blooms at the same sampling site often showed similar bacterioplankton community members responding to these blooms (Needham et al., 2018, Needham & Fuhrman, 2016).

Other long-term ecological research stations that regularly experience phytoplankton blooms are, for example, the coastal observatories at the North Sea island of Helgoland, the L4 station at Plymouth in the English Channel, the Linnaeus Microbial Observatory in the Baltic Sea, the Delaware estuary and the Blanes Bay in the Mediterranean Sea. At all of these coastal sites and in other studies investigating only a single phytoplankton bloom, secondary blooms of heterotrophic bacterioplankton are often dominated by the same bacterial clades. These are predominantly marine members of the *Bacteroidetes*, *Gammaproteobacteria* and the family *Rhodobacteraceae* of the *Alphaproteobacteria* (Gilbert et al., 2012, Landa et al., 2016, Lindh et al., 2015, Needham & Fuhrman, 2016, Taylor et al., 2014, Teeling et al., 2012, Williams et al., 2013, Yang et al., 2015). Members of these clades are specialized on uptake of phytoplankton-derived OM, but individual strategies differ. *Rhodobacteraceae* can take up a rather broad substrate spectrum including dimethylsulfoniopropionate, urea, phosphoesters or phosphonates (reviewed in Buchan et al. (2014)), and possess transporters of the tripartite ATP-independent periplasmic (TRAP), major facilitator superfamily (MFS) and ATP-binding cassette (ABC) families (Moran et al., 2007, Newton et al., 2010, Rinta-Kanto et al., 2012, Teeling et al., 2012). *Bacteroidetes* and *Gammaproteobacteria* have transporter profiles often dominated by transporters of the TonB-dependent transporter (TBDT) systems that are regularly upregulated during blooms (Klindworth et al., 2014, Rinta-Kanto et al., 2012, Tang et al., 2012, Teeling et al., 2012, Williams et al., 2013). The TBDT system allows for uptake of molecules larger than 600 Da, and thus permits the uptake of HMW organic molecules including small oligomers. In addition to the different transporter profiles, *Bacteroidetes* and also *Gammaproteobacteria* often possess a larger number of hydrolytic enzymes for the degradation of DOM (Barbeyron et al., 2016b, Fernández-Gómez et al., 2013, Xing et al., 2015).

Prominent clades from the *Bacteroidetes* that have been frequently found to respond to phytoplankton blooms are *Formosa*, *Polaribacter* and *Ulvibacter* species, the *Cryomorphaceae* family or the NS3a and NS5 marine groups (Chafee et al., 2017, Lindh et al., 2015, Needham et al., 2018, Needham & Fuhrman, 2016, Tan et al., 2015, Teeling et al., 2012, Williams et al., 2013, Yang et al., 2015). Even though it is often the same genera responding to phytoplankton blooms, their individual contributions vary and they typically show a successional response pattern with variation in the order of genera (Chafee et al., 2017, Teeling et al., 2012). Often *Bacteroidetes* respond with a rapid increase in relative abundance, which can be delayed compared to the onset of the phytoplankton bloom. During phytoplankton bloom events, bacteria can occur either in close connection with cell-to-cell contact, inhabiting the phytoplanktons' phycosphere, or occur as free-living organisms. From an abundance perspective, the free-living fraction makes up the majority of the bacterial cells (e.g. Ghiglione et al. (2009), Li et al. (2015b),

Turley & Stutt (2000)) and is thus the major fraction involved in the remineralization of algae-derived DOM.

By now there have been many studies focusing on the overall response patterns of larger groups of bacterioplankton to phytoplankton bloom events. These have mostly focused on genus level response patterns using 16S rRNA sequence analyses. More in depth studies that focus on the ecological roles of individual species or subspecies have only recently emerged (e.g. Chafee et al. (2017), Delmont et al. (2015), Hugerth et al. (2015)). Further studies would be an important step to shed light on individual species niche specialization during algae blooms and could help to elucidate the degree of niche specialization among closely related taxa.

1.4 Marine polysaccharides

Polysaccharides can contribute a major fraction of DOM and POM. They are carbon-rich compounds built from monosaccharides that are linked via glycosidic bonds. They consist mostly of carbon, hydrogen and oxygen atoms with a ratio of $C_n(H_2O)_n$, and can sometimes be decorated with additional heteroatoms, such as nitrogen, phosphorus and sulphur. In oligo- and polysaccharides, monosaccharides are linked via α- or β-glycosidic linkages, depending on the first monomer's stereochemical configuration at the anomeric centre of the first carbon atom. Linkages can occur between the anomeric carbon of one monosaccharide and any hydroxyl group of a second monosaccharide. Based on these linkage positions the linkage is termed for example "1,3". This means that the glycosidic bond is formed between the C1 atom of the first monomer and the C3 atom of the second monomer. Polysaccharides can either be linear or show different degrees of branching patterns, and can be linked at different carbon positions. Furthermore, they can be homopolysaccharides consisting of only one monosaccharide type, e.g. starch, or heteropolysaccharides consisting of different monosaccharides, such as galactomannans. All these factors support a theoretically very high diversity of polysaccharides (Laine, 1994).

In macroalgae polysaccharides make up about 50% of their biomass (Kloareg & Quatrano, 1988) and occur as storage or cell wall components. Macroalgae are predominantly found in coastal ecosystems, where they make up a large amount of the primary biomass (Gobet et al., 2018). Major groups of macroalgae are green algae (*Chlorophyta*) and red algae (*Rhodophyta*) from the *Archaeplastida* that both also include unicellular algae. Brown algae (*Phaeophyceae*) are closely related to diatoms (*Bacillariophyceae*) and both belong to the *Stramenopiles*. While all these three groups contain cellulose as a crystalline polysaccharide in their cell wall, red algae typically contain agars and

carrageenans, green algae contain ulvans, and brown algae alginates or fucose-containing sulphated polysaccharides (FCSP) as matrix polysaccharides (reviewed in Popper et al. (2011)). The major storage polysaccharides for macroalgae are starch in red and green algae and laminarin in brown algae (Gobet et al., 2018).

Polysaccharides in microalgae (phytoplankton) are present as storage compounds, cell wall components and exudates that are released into the cell surroundings. Overall polysaccharide contribution to the total cell volume can vary between 13-90% (Myklestad, 1974). Structures of polysaccharides derived from phytoplankters are less well studied compared to macroalgal polysaccharides and remain to be resolved for many phytoplankton species. Nevertheless, monosaccharide composition of phytoplankters are often dominated by a handful of monomers. Among those are glucose, mannose, fucose, arabinose, xylose, rhamnose and galactose (reviewed in Gügi et al. (2015) and Mühlenbruch et al. (2018)). These monosaccharide compositions vary not only depending on phytoplankton species but also depending on the life cycle of the phytoplankter and on the cellular localization of these polysaccharides (Chiovitti et al., 2003, Gügi et al., 2015, Urbani et al., 2005). Within diatoms and other phytoplankters glucose is for example the dominating monomer in storage compounds (Gügi et al., 2015, Myklestad & Granum, 2009), while mannose was detected as the major cell wall monosaccharide in many diatom species (Gügi et al., 2015). Recently the polysaccharide structure of a sulphated glucuronomannan has been resolved from the diatom species *Phaeodactylum tricornutum* (Le Costaouëc et al., 2017). This sulphated glucuronomannan was shown to have an α-1,3-mannan backbone decorated with glucuronic acid and sulphate ester groups (Le Costaouëc et al., 2017). The major storage polysaccharide in diatoms is chrysolaminarin (Beattie et al., 1961), which usually consists of 20 to 30 glucose monomers, and is characterized by a β-1,3-glucose backbone with occasional β-1,6-glucose branches. But even in this quite simple major storage polysaccharide there is a high diversity, including different degrees of polymerization, varying degree of branching, and sometimes β-1,2-glucose side chains (Garcia-Vaquero et al., 2017, Gügi et al., 2015).

This high polysaccharide diversity and the often unknown structures make it difficult to understand the overall polysaccharide profile present during phytoplankton blooms. However, by studying the monosaccharide compositions of the total combined carbohydrates, Sperling et al. (2017) could show that glucose and co-eluting mannose and xylose were the dominating monosaccharides during a phytoplankton spring bloom at the island of Helgoland in the German North Sea. Other tools used to infer polysaccharide composition focus on the degradation potential of the bacterial community to elucidate which polysaccharides the bacteria are adapted to. For example, bacterial laminarinase activities have been shown to be abundant in ocean surface waters (Arnosti, 2011, D'Ambrosio

et al., 2014, Reintjes et al., 2018), which suggest a high prevalence of the polysaccharide laminarin. This has recently been confirmed by a newly developed method that uses an enzyme assay to enzymatically quantify laminarin concentration in marine environmental samples (Becker et al., 2017). Enzymes for this assay were derived from a laminarin degrading genetic locus and revealed concentrations of laminarin of 0.48 ± 0.09 mg/L for the 10-µm POM fraction and 0.13 ± 0.02 mg/L for the 3-µm POM fraction during a phytoplankton bloom at the North Sea island of Helgoland (Becker et al., 2017). These methods depend, however, on a priori knowledge about the actual polysaccharides and the enzymes involved in their degradation.

1.5 Polysaccharide utilization by *Bacteroidetes*

Bacteroidetes are well adapted for the uptake of HMW DOM and in particular the degradation of polysaccharides. Their genomes are characterized by gene clusters that are involved in the degradation of polysaccharides, termed polysaccharide utilization loci (PULs) (Bjursell et al., 2006). The archetypical polysaccharide utilization locus (PUL) is characterized by a gene tandem of a SusC-like TBDT and a SusD-like protein. These terms (SusC and SusD) refer to proteins and their corresponding genes (*susC* and *susD*) described in the first PUL in *Bacteroides thetaiotaomicron* (Reeves et al., 1997) that are part of the starch utilization system (Sus). Genes of this PUL are involved in the degradation (e.g. *susA*, *susB*, *susG*), the binding (e.g. *susD*, *susE* and *susF*) and the uptake of starch into the periplasm (*susC*) and have been extensively studied in this model PUL (reviewed in Hemsworth et al. (2016)). While similar gene clusters, also harbouring a TBDT, are present in other bacterial phyla such as the *Alteromonadales* (Gobet et al., 2018, Koch et al., 2018, Neumann et al., 2015) the unique feature of the *Bacteroidetes* is the SusD-like protein (Grondin et al., 2017). Together with the TBDT the SusD-like protein forms a pedal bin structure (Glenwright et al., 2017), in which the SusD protein acts as a lid that closes upon substrate binding. At the same time the SusC-like TBDT changes its conformation and opens a channel into the periplasmic space, allowing for substrate uptake (Glenwright et al., 2017).

In order to completely degrade one polysaccharide substrate, bacteria need a set of enzymes that are able to degrade the different types of linkages present in the respective polysaccharide and eventually take up the cleavage products into the periplasm and finally the cytoplasm. This set of enzymes includes the SusC- and SusD-like proteins, a set of degradative carbohydrate-active enzymes (CAZymes) and some additional proteins such as sulphatases, proteases or response regulators. CAZymes can be involved in the degradation, modification or biosynthesis of polysaccharides and are separated into

five enzyme classes in the CAZy database (Lombard et al., 2014). These are degradative CAZymes of the glycoside hydrolase (GH), polysaccharide lyase (PL), and carbohydrate esterase (CE) families as well as biosynthetic enzymes of the glycosyltransferase (GT) families and auxiliary activities (AAs). AAs include redox enzymes acting together with CAZymes (Lombard et al., 2014). Additionally, there are associated modules involved in the binding of polysaccharides, termed carbohydrate-binding modules (CBMs) (Lombard et al., 2014). Each of these enzyme classes contains several families that are defined by amino acid sequence similarity. Thus some of these families are broad in their functional spectrum, and can include diverse functions, while others contain enzymes with similar or so far only one described function. Some of these families have been further divided into subfamilies. Among these families are the GH5 (Aspeborg et al., 2012), GH13 (Stam et al., 2006), GH30 (St John et al., 2010), GH43 (Mewis et al., 2016) and the PL families (Lombard et al., 2010). By analysing the different CAZyme families present on PULs, and considering their individual functions, the structures or at least the linkage types present in their target polysaccharides can be predicted.

The advantage of a PUL-like uptake system is substrate hoarding. The initial cleavage of polysaccharides into oligosaccharides by extracellular CAZymes allows for quick uptake of the oligosaccharides into the periplasm of the bacterial cell via the TBDT (Fig. 1.3). In the periplasm these oligosaccharides are protected from competing bacteria and can be further degraded into monosaccharides that can pass via dedicated transporters into the cytoplasm. Depending on the efficiency of uptake of extracellular cleavage products, this mechanism can either be selfish with all cleavage products taken up, as described for α-mannan degradation by a human gut *Bacteroidetes* (Cuskin et al., 2015) or semi-selfish with some cleavage product left for other bacteria (Rakoff-Nahoum et al., 2016, Reintjes et al., 2018). Uptake of fluorescently labelled polysaccharides into the periplasmic space has already been shown for marine *Bacteroidetes*, suggesting a selfish or semi-selfish uptake mechanism (Reintjes et al., 2018, 2017).

The study of PULs was initially dominated by the research field studying human gut microbiota, as evident from the first version of the PUL database (PUL-DB, Terrapon et al. (2015)). In the beginning PUL-DB included genomes of 67 *Bacteroidetes* species from the gastrointestinal tract and only two other *Bacteroidetes* species, one of which was *Flavobacterium johnsoniae* that is commonly found in soil and freshwater (Bernardet et al., 1996, Stanier, 1947). By now the PUL-DB, a database describing experimentally characterized and predicted PULs from *Bacteroidetes*, includes PUL predictions for 820 species from various environments (Terrapon et al., 2018). Within the last decades a number of PULs of marine *Bacteroidetes* have been studied in detail, and their substrate specificity and degradation pathway has been experimentally verified. Among these are, for example, PULs that target agar and porphyran in a human gut *Bacteroidetes*, which

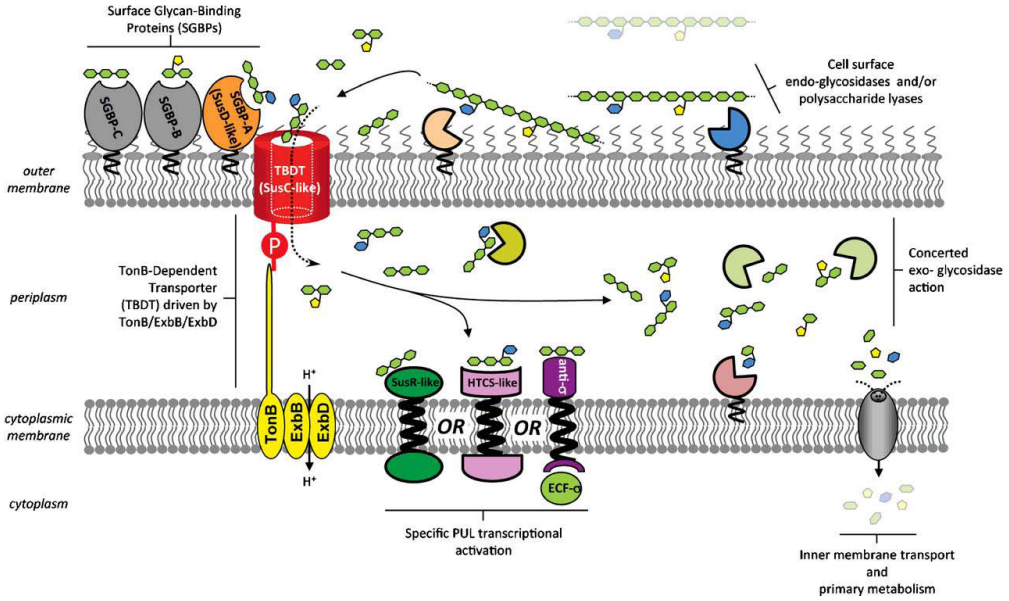


Figure 1.3: Cellular arrangement of PUL-encoded components for glycan processing in Gram-negative bacteria, based on the archetypal *Bacteroidetes* Sus. The number of cell-surface glycan-binding proteins (SGBPs) and GHs vary among PULs, as does the presence and nature of carbohydrate sensor/transcriptional regulator proteins (SusR-like, HTCS-like or anti-σ/ECF-σ) (Hemsworth et al., 2016).

were likely obtained through horizontal gene transfer from a marine bacterium (Hehemann et al., 2010). PULs targeting alginate were described in *Zobellia galactanivorans* Dsij^{T} (Thomas et al., 2012) and *Gramella forsetii* KT0803 (Kabisch et al., 2014). Laminarin as a target substrate was verified for PULs in *Gramella forsetii* KT0803 (Kabisch et al., 2014) and *Polaribacter* species (Xing et al., 2015). A carrageenan PUL was verified for *Zobellia galactanivorans* Dsij^{T} (Ficko-Blean et al., 2017) and an ulvan PUL for *Formosa agariphila* KMM 3901^{T} (Reisky et al., 2018, Salinas & French, 2017). Additional studies have focused on PULs in *Gramella flava* JLT2011 (Tang et al., 2017), PULs in different *Cellulophaga* species (Valdehuesa et al., 2018) and the overall PUL repertoire present in *Zobellia galactanivorans* Dsij^{T} (Barbeyron et al., 2016b).

Studies with bacterial isolates are very important for the understanding of individual PUL functions and the relevance of individual enzymes involved in the degradation of polysaccharides. The disadvantage of these studies is that the isolates are often not representative of the natural marine microbial community. Studies focusing on the environmentally relevant marine PULs are so far limited (e.g. Gómez-Pereira et al. (2012)). Recently, PUL analyses in metagenomic studies that investigate the microbial community without prior need for isolation, have started to emerge. So far these studies focused on the fecal microbiome of the North American beaver (Armstrong et al., 2018), the gut microbiota of a wood-feeding higher termite (Liu et al., 2018) and the moose rumen (Svartström et al., 2017). The application of metagenomic analyses to study

PULs in marine microbes is an important step towards extending the current knowledge of marine PULs and to reveal the most relevant PULs of the marine *Bacteroidetes* community.

1.6 Metagenome analyses - recent technological and methodological advances

Metagenomics is the direct cultivation-independent sequencing of natural microbial communities by extracting whole communities' genomic information and studying their genomic potential. The field of metagenomics began with the cloning and subsequent sequencing of larger genomic DNA fragments of about 30 to 100 kb that were cloned into fosmids or bacterial artificial chromosomes (Béjà et al., 2000b, Stein et al., 1996). These fragments were screened for phylogenetic markers (e.g. the 16S rRNA genes) and could help to shed light on parts of the individual organisms' functions. The large insert size approaches have led to the discovery of proteorhodopsin (Béjà et al., 2000a) and allowed for studying PULs in uncultured marine *Bacteroidetes* communities (Gómez-Pereira et al., 2012). Further developments in sequencing technologies in the early 2000s led towards the development of short read high-throughput sequencing (HTS). HTS was much more cost-efficient. It did not require cloning and allowed for metagenomic sequencing of larger fractions of the microbial community using short DNA fragments.

1.6.1 Sequencing technologies

Until the start of HTS in the early 2000s the standard sequencing technology since the late 1970s was Sanger sequencing (Sanger et al., 1977). The Sanger sequencing method produces relatively long (up to 1000 bp), high quality sequences, but it is costly and of low throughput (Vincent et al., 2017). Thus it is laborious to use Sanger sequencing even for small genomes. When 454 Life Sciences introduced the first HTS technology in the early 2000s it allowed scientist to upscale their sequencing efforts and sequence millions of DNA molecules in a single run. Since then sequencing technologies have advanced quickly. Currently the most commonly used sequencing platforms for metagenomic HTS are the Illumina platforms. These can reach maximum output of larger than 1000 Gb per run. Drawbacks are primarily related to the limited read length of only up to 300 bp. Sequencing technologies producing longer reads of larger 10 kb are, for example, the SMRT technology from Pacific Biosciences as implemented in their Sequel instrument (McCarthy, 2010) or the nanopore sequencing technology as implemented in the MinION instrument from Oxford Nanopore (Mikheyev & Tin, 2014). However,

these technologies have higher error rates, a rather low throughput and require high amounts of input DNA, which for environmental samples might not always be feasible.

1.6.2 Assembly strategies

With the development of new HTS technologies came the necessity to develop strategies that allow for the assembly of the sequenced reads into longer and more informative stretches of DNA. Early assembly strategies for long Sanger reads could be based on overlap-layout consensus approaches (reviewed in Simpson & Pop (2015), Vollmers et al. (2017)). Current HTS, on the other hand, requires assembly strategies without an all to all comparison of reads, as this is computationally impractical with billions of short reads in a single dataset (current maximum output of an Illumina HiSeq 4000 sequencer: 10 Billion paired-end reads per run). In the overlap-layout consensus approach each read is compared to all other reads of the dataset to detect overlaps of all reads (Fig. 1.4). Then a graph is constructed, with each read represented by a node in the graph and overlapping reads linked by an edge. Each node in this graph can for simplicity be seen as a location on a map, with different locations linked by pathways (edges) (Vollmers et al., 2017). From this data structure, an appropriate layout of the reads is extracted by finding the best path through the graph, with which each element is visited in the correct order forming a continuous stretch of sequence, so called contigs. The consensus sequence for a contig is then computed by considering the consensus nucleotide for each position, which is represented by the majority of reads at this position (Vollmers et al., 2017).

The de Bruijn graph assembly, which at the moment is the most commonly used heuristic in metagenome assemblers, has the advantage that it does not require the computationally time consuming step of finding overlaps between all pairs of reads. Instead in a de Bruijn graph assembly all reads are broken into a sequence of overlapping k-mers, i.e. a subsequence of length k. Each neighbouring k-mer will be identical with the preceeding and succeeding k-mer on a read in k-1 of their bases. In the graph these overlapping parts of the k-mers (k-1mers) are represented by a node with the connections between these overlaps (edges) represented by the k-mers itself (Fig. 1.5). In this way overlaps between different reads are directly incorporated into the graph as it is constructed from the k-mers of all reads (Fig. 1.5 C). Assemblers will then try to construct contigs from this data structure by finding a path through the graph that visits each edge in the graph only once (reviewed in Simpson & Pop (2015), Vollmers et al. (2017)). The most commonly used metagenomic assemblers that use the de Bruijn graph assembly method are MetaSPAdes (Nurk et al., 2017), MEGAHIT (Li et al., 2015a) and IDBA-UD (Peng et al., 2012) that all use iterative multi-k-mer approaches and have been

shown to perform well on natural microbial samples (Vollmers et al., 2017). After the assembly, functions of individual genes or gene clusters can be predicted using, for example, homology- or sequence motif-based comparisons to closely related proteins with known functions.

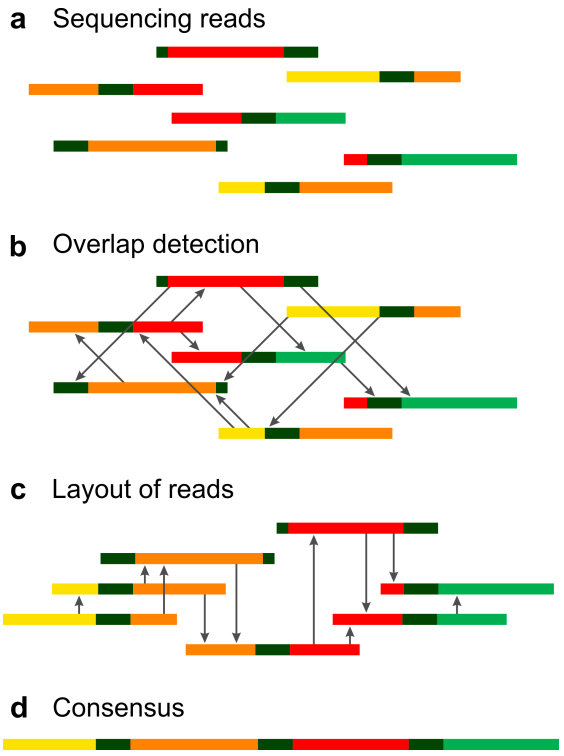


Figure 1.4: Schematic overview of the traditional overlap layout consensus (OLC) assembly approach. Different colors symbolize sequence stretches from different genomic regions. Grey stretches indicate small repetitive regions. (A) Random shotgun sequencing reads are obtained. (B) The read ends of all sequences are aligned to each other, in order to detect overlaps. (C) Even though all reads contain repetitive regions, the reads can be unambiguously placed within the layout graph, based on unambiguous overlaps at the read ends. (Vollmers et al. (2017); licensed under CC BY 4.0)

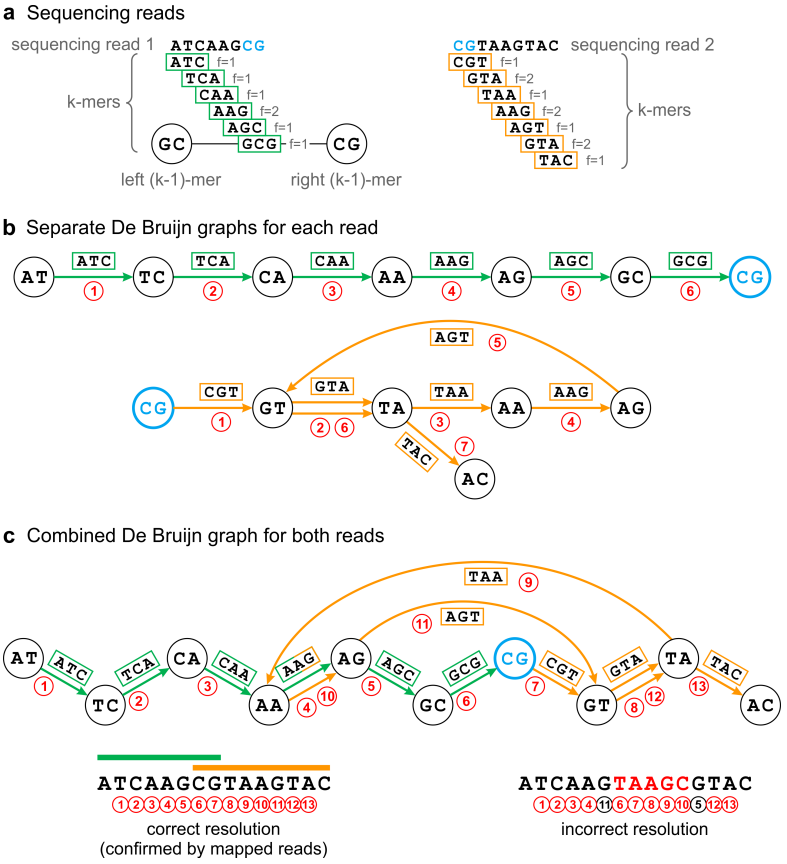


Figure 1.5: Schematic overview of the de Bruijn graph concept. For illustration purposes, the sequencing data is represented by only two reads, read1 and read2, which overlap by only two basepairs (marked in blue). The k-mer length is set to k=3. (A) All overlapping k-mers are derived from the reads and counted (f = frequency). Neighboring k-mers are shifted by one basepair. This means adjacent k-mers share exactly k-1 basepairs (k-1mers) to their left or right. On the other hand, this also means that neighbouring k-mers may be identified, based on identical left or right k-1mers. (B) Examples of de Bruijn graphs constructed separately for each read. Each k-mer represents an edge connecting its left and right k-1mers (nodes). K-mers which are identical in their left or right k-1mer share a node in the graph, indicating that they may occur adjacent to each other. By following these connections, it is possible to reconstruct the original read sequences. Graphs are resolved by following these paths beginning at nodes with less ingoing than outgoing edges (indicating the start of a path) and ending at nodes with less outgoing than ingoing edges (indicating the end of a path). The graph for read1 (green) forms a single straight path (visiting edges 1-6 consecutively) because all k-mers are unique within this read (although k-mer "AAG" also occurs on read2). In contrast, read2 (orange) contains a repetitive k-mer ("GTA", edges 2+6) which introduces a branched path to the graph. Nevertheless, there exists only one possible path that visits every node exactly once (visiting edges 1-7 consecutively), illustrating that repetitive regions can be resolved without breaking the de Bruijn graph. (C) Combined de Bruijn graph for read1 (green) and read2 (orange). An additional branch is introduced to the graph by the repetitive k-mer "AAG" occurring in both reads. As a result, the graph may be equally resolved by following different paths. However, the correct path can be identified by mapping the reads back to the graph. (Vollmers et al. (2017); licensed under CC BY 4.0)

1.6.3 Metagenome binning and taxonomic classification of metagenome-assembled genomes

Linking functions, predicted by annotation of individual metagenomic contigs, to discrete microbial species is required to predict its ecological niche. This can be achieved by binning of metagenomic contigs into so-called metagenome-assembled genomes (MAGs; Hugerth et al. (2015)), a common approach for further processing of metagenomic datasets. Binning algorithms use a combination of the sequence composition and coverage information of contigs to group them into MAGs. Sequence composition of contigs can be the GC content or the tetranucleotide frequency pattern of contigs. Coverage of contigs is calculated by mapping the metagenomic reads onto the contigs and extracting abundance information for each individual contig. Here a time-series dataset is of high value as the abundance of a single contig can be tracked over time, which helps for the binning process, as contigs originating from the same organism should show similar abundance patterns over time whereas those from different organisms should split. Common binning programs are CONCOCT (Alneberg et al., 2014), GroopM (Imelfort et al., 2014), MaxBin (Wu et al., 2014) or MetaBAT (Kang et al., 2015). After binning the quality of MAGs should be assessed by using tools (e.g. CheckM; Parks et al. (2015)) that consider the presence and absence of single-copy marker genes. Single-copy marker genes are genes that should only occur as a single copy in an organisms' genome and are thus good estimators of genome completeness and contamination.

High quality MAGs can help to broaden our knowledge about species diversity and metabolic processes in the marine environment. Many of these MAGs are lacking closely related genomes in public databases and thus are likely representing new species or even new phyla (e.g. Hug et al. (2016)). Instead of defining these species boundaries by the classical 16S rRNA-based taxonomy, a new taxonomic approach has become more commonly used in the last years that is based on genome phylogeny using a set of concatenated proteins. Tools like CheckM (Parks et al., 2015) or the GTDB$_{tk}$ (Parks et al., 2018) place MAGs in reference genomes trees to evaluate their taxonomic relationships. GTDB$_{tk}$ is based on the genome taxonomy database (GTDB), which improved genome taxonomy by conservatively removing polyphyletic groups and normalizing taxonomic ranks on the basis of relative evolutionary divergence (Parks et al., 2018). In addition to genome placement of MAGs in reference genome trees, direct measures of genome relatedness are important to evaluate a MAG's novelty. This can be achieved by calculating the average nucleotide identity (ANI) between two genomes. An ANI value of ∼95% is commonly used as a species boundary (Goris et al., 2007, Konstantinidis & Tiedje, 2005). This novel definition of microbial species will in the future become more important, as more and more metagenomic studies focus on the extraction of MAGs (Brown

et al., 2015, Delmont et al., 2015, 2018, Hugerth et al., 2015, Parks et al., 2017) and will lead to the increase of MAGs in public databases (e.g. Bowers et al. (2017)). Thus there is also the need to validly describe these genomes of uncultured organism as new *Candidatus* species (Konstantinidis & Rossello-Mora, 2015, Konstantinidis et al., 2017). Metagenomic binning followed by functional annotation and taxonomy will thus help to get a more holistic view on the marine microbial community. In particular, it would enable in-depth studies of so far uncultivable bacteria with a focus on polysaccharide degradation.

1.7 Aims of the study

The overall goal of the work presented here was to shed light on the PUL repertoires of marine *Bacteroidetes* and to expand the knowledge about their role in polysaccharide degradation. The major body of work in this thesis was focused on the *Bacteroidetes* occurring during spring phytoplankton blooms at the coastal research station Helgoland, an island in the North Sea (Fig. 1.2). Research at the long-term research site "Kabeltonne" has been conducted since 1962 and includes the collection of physicochemical parameters (e.g. temperature, salinity, nutrients, chlorophyll *a*) and microbiological data, such as total bacterial cell and phytoplankton counts (Wiltshire et al., 2010). The data from Helgoland spring phytoplankton blooms presented in this thesis stems from samples collected from the 0.2-3 µm size fraction of the bacterioplankton in the years 2009 to 2012 with a sampling resolution of about one to two times a week. This thesis reports on the following four work packages:

i) The evaluation of recurrent patterns of individual bacterioplankton clades and specific individual species during phytoplankton blooms was addressed in Chapter 2 and Chapter 4, respectively. In Chapter 2 the analysis was focused on recurrent taxonomic and functional patterns of major bacterioplankton clades analysed using the 16S rRNA gene and unbinned, but taxonomically classified metagenomes. In Chapter 4 this question was addressed with a higher sampling resolution of metagenomes and the focus on individual *Bacteroidetes* species derived through metagenome binning.

ii) A more detailed analysis of the individual *Bacteroidetes* species degradation potentials and their PUL repertoires was addressed in Chapter 3 and Chapter 4. Chapter 3 focused on the PUL content of 53 *Bacteroidetes* isolates that were isolated from different locations in the North Sea. Chapter 4 focused on PULs within MAGs of the most abundant *Bacteroidetes* species during the 2010 to 2012 phytoplankton blooms and shed light on the most relevant PULs present in the free-living *Bacteroidetes* community.

iii) Differences of PUL repertoires of the *Bacteroidetes* community at two contrasting open ocean sites in the North Atlantic were investigated in Appendix 1 using *Bacteroidetes*-affiliated fosmids.

iv) Appendix 2 and Appendix 3 addressed individual *Bacteroidetes* species and their ecological roles during phytoplankton spring blooms at Helgoland. In Appendix 2 two isolates of the *Formosa* genus were studied in detail and in Appendix 3 MAGs from the *Ulvibacter* genus were taxonomically and functionally described and reclassified as *Candidatus* Prosiliicoccus species.

Chapter 2

Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms

Hanno Teeling[1], Bernhard M Fuchs[1], Christin M Bennke[1], Karen Krüger[1], Meghan Chafee[1], Lennart Kappelmann[1], Greta Reintjes[1], Jost Waldmann[1], Christian Quast[1], Frank Oliver Glöckner[1], Judith Lucas[2], Antje Wichels[2], Gunnar Gerdts[2], Karen H Wiltshire[2] and Rudolf I Amann[1]

[1]Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany
[2]Alfred-Wegener-Institute Helmholtz-Center for Polar and Marine Research, Biological Station Helgoland, Kurpromenade 201, 27498 Helgoland, Germany

Published in *eLife* (DOI: 10.7554/eLife.11888).

Contributions to the manuscript:
Experimental concept and design: 5%
Acquisition of experimental data: 0%
Data analysis and interpretation: 35%
Preparation of figures and tables: 30%
Drafting of the manuscript: 15%

Electronic figure versions and supplementary material are available at https://doi.org/10.7554/eLife.11888.

2.1 Abstract

A process of global importance in carbon cycling is the remineralization of algae biomass by heterotrophic bacteria, most notably during massive marine algae blooms. Such blooms can trigger secondary blooms of planktonic bacteria that consist of swift successions of distinct bacterial clades, most prominently members of the *Flavobacteriia*, *Gammaproteobacteria* and the alphaproteobacterial *Roseobacter* clade. We investigated such successions during spring phytoplankton blooms in the southern North Sea (German Bight) for four consecutive years. Dense sampling and high-resolution taxonomic analyses allowed the detection of recurring patterns down to the genus level. Metagenome analyses also revealed recurrent patterns at the functional level, in particular with respect to algal polysaccharide degradation genes. We, therefore, hypothesize that even though there is substantial inter-annual variation between spring phytoplankton blooms, the accompanying succession of bacterial clades is largely governed by deterministic principles such as substrate-induced forcing.

2.2 eLife digest

Small algae in the world's oceans remove about as much carbon dioxide from the atmosphere as land plants. These algae do not grow continuously, but often surge in numbers during temporary blooms. Such blooms can be large enough to be seen from space by satellites. The lifespan of algae within such blooms is short, and when they die, marine bacteria feed on the remnants, which releases much of the stored carbon dioxide.

Much of an algal cell consists of different types of polysaccharides. These large molecules are essentially made from sugars linked together. Polysaccharides are varied molecules and can contain many different sugars that can be linked in a number of different ways. During algae blooms bacteria proliferate that are specialized in the degradation of these polysaccharides. In 2012, researchers reported how over the progression of an algae bloom different groups of marine bacteria bloomed in rapid succession. However, it remained unknown whether the same or different groups of bacteria respond to algae blooms at the same place from year to year, and whether or not these bacteria use the same enzymes to degrade the polysaccharides.

Teeling, Fuchs et al. – who include many of the researchers from the 2012 study – now report on the analysis of a series of algae blooms that occurred in the southern North Sea between 2009 and 2012. The analysis is based on samples collected every week during the spring seasons, and shows that certain groups of related bacteria, known as clades, became common during each bloom. Teeling, Fuchs et al. also found indications

that the clades that repeatedly occurred had similar sets of genes for degrading algal polysaccharides, but that the sets were different between the clades.

These data suggest that there is a specialized bacterial community that together can degrade the complex mixture of algal polysaccharides during blooms. This community reappears each year with an unexpectedly low level of variation. Since different species of algae made up the blooms in each year, this finding suggests that the major polysaccharides in these algae are similar or even identical.

Future work will focus on the specific activities of bacterial enzymes that are needed to degrade polysaccharides during algae blooms. Study of these enzymes in the laboratory will help to resolve, which polysaccharides are attacked in which manner, and to ultimately help to identify the most abundant algal polysaccharides. This will improve our current understanding of the carbon cycle in the world's oceans.

2.3 Introduction

Pelagic zones of the world's oceans seemingly constitute rather homogeneous habitats, however, they feature enough spatial and temporal variation to support a large number of species with distinct niches. This phenomenon has been termed "paradox of the plankton" by G. Evelyn Hutchinson (Hutchinson, 1961). Interactions within planktonic microbial communities are manifold and complex (see Amin et al. (2012) and Worden et al. (2015) for reviews). Still, planktonic microbial communities are simple in comparison to benthic or terrestrial soil communities and thus particularly suitable for the study of microbial community composition dynamics. In recent years, continuous biodiversity studies at long-term sampling stations have started to reveal discernible deterministic patterns within marine microbial plankton communities (see Fuhrman et al. (2015) for a recent review). This is particularly true for less dynamic oligotrophic oceanic regions that are dominated by the members of the alphaproteobacterial *Pelagibacteriaceae* (SAR11 clade) and the cyanobacterial *Prochlorococcaceae* (*Prochlorococcus marinus*). By contrast, more dynamic eutrophic coastal regions are subject to frequent system perturbations and thus seldom in a state of equilibrium. This can lead to apparently stochastic changes in bacterioplankton community composition. To capture recurrence of biodiversity patterns in such coastal areas, sampling must occur at the order of weekly to sub-weekly time scales over multiple years. Owing to the lack of such intensively sampled long-term time series data, our current understanding of the extent and predictability of recurring microbial biodiversity patterns for such marine habitats is still limited.

A particularly important connection in the marine carbon cycle exists between marine microalgae as primary producers and heterotrophic bacteria that feed on algal biomass. Global photosynthetic carbon fixation is estimated to exceed 100 Gigatons yearly, of which marine algae contribute about half (Falkowski et al., 1998, Field et al., 1998, Sarmento & Gasol, 2012). Planktonic uni- to pluricellular algae such as diatoms, haptophytes, and autotrophic dinoflagellates are the most important marine primary producers. Diatoms alone are estimated to contribute 20–40% to global carbon fixation (Armbrust, 2009, Mann, 1999, Nelson et al., 1995).

Primary production by planktonic microalgae differs from primary production by sessile macroalgae or land plants as it is much less constant, but culminates in blooms that are often massive, as occurs worldwide during spring blooms from temperate to polar regions. These blooms are highly dynamic phenomena that are time-limited by nutrients, predator grazing and viral infections. Bloom termination results in a short-lived massive release of algal organic matter that is consumed by dedicated clades of heterotrophic bacterioplankton. This trophic connection leads to synchronized blooms of planktonic bacteria during phytoplankton blooms, as has been described in various studies (Bell

& Kuparinen, 1984, Niu et al., 2011, Tada et al., 2011, Tan et al., 2015, Teeling et al., 2012, Yang et al., 2015).

The activities of these heterotrophic bacteria impact the proportion of algal biomass that is directly mineralized and released back into the atmosphere mostly as carbon dioxide, and the algae-derived biomass that sinks out to the bottom of the sea as carbonaceous particles. These are further remineralized by particle-associated bacteria while sinking and by benthic bacteria when reaching the sediment, even in the deep sea (e.g., (Ruff et al., 2014)). The remainder is buried for a long time as kerogen and forms the basis for future oil and gas reservoirs. The ratio between bacterial mineralization and burial of algae-derived organic matter thus has a profound influence on the atmospheric carbon dioxide concentration (Falkowski et al., 1998). However, the bulk of bacteria during phytoplankton blooms are free-living and not attached to particles or algae. These bacteria play a pivotal role in the mineralization of algae-derived non-particulate dissolved organic matter (DOM).

The bacterial clades that respond most to phytoplankton blooms belong to the classes *Flavobacteriia* (phylum *Bacteroidetes*) and *Gammaproteobacteria*, and the *Roseobacter* clade within class *Alphaproteobacteria* (Buchan et al., 2014). This response is typically not uniform, but consists of a series of distinct clades that bloom one after another. In the year 2009, we investigated the response of bacterioplankton to a diatom-dominated spring phytoplankton bloom in the German Bight (Teeling et al., 2012). Within the free-living bacteria (0.2 to 3 μm) we observed a swift succession of bacterial clades that were dominated by *Flavobacteriia* and *Gammaproteobacteria*, with consecutively blooming *Ulvibacter* (*Flavobacteriia*), *Formosa* (*Flavobacteriia*), *Reinekea* (*Gammaproteobacteria*), *Polaribacter* (*Flavobacteriia*) genera and SAR92 (*Gammaproteobacteria*) as prominent clades.

Using time-series metagenome and metaproteome analyses, we demonstrated that the substrate-spectra of some of these clades were notably distinct. The succession of bacterioplankton clades hence constituted a succession of distinct gene function repertoires, which suggests that changes in substrate availability over the course of the bloom were among the forces that shaped the bacterioplankton community. Dominance of bottom-up over top-down control is assumed to be characteristic for the initial phases of spring phytoplankton blooms. After winter, inorganic nutrients are aplenty, and the overall abundance of microbes is low. When suitable temperature and sunlight conditions are met in spring, algae and subsequently bacteria can enter an almost unrestricted proliferation. In contrast, predators such as flagellates, protists and zooplankton can only start proliferating when their food sources are available in larger numbers. Hence, top-down

control by predation sets in only during later bloom phases. This situation is distinct from summer and fall phytoplankton blooms.

Pronounced differences between blooming clades were found in the gene frequencies and protein expression profiles of transporters and carbohydrate-active enzymes (CAZymes; (Cantarel et al., 2009, Lombard et al., 2014)), such as glycoside hydrolase (GH), polysaccharide lyase (PL), carbohydrate esterase (CE), or carbohydrate-binding module (CBM) containing genes. The latter indicates a pronounced niche partitioning with respect to algal polysaccharide degradation. Marine algae produce large quantities of distinct polysaccharides, for example storage, cell matrix and cell wall constituents, or as part of extracellular transparent exopolymer particles (TEP). It has been recently shown that in particular *Flavobacteriales* and *Rhodobacterales* respond to TEP availability (Taylor et al., 2014). The diversity of algal polysaccharides is too high for a single bacterial species to harbor all the genes required for the complete degradation of all naturally occurring variants. Thus, polysaccharide-degrading bacteria specialize on dedicated subsets of polysaccharides, which is why the decomposition of algal polysaccharides during and after algal blooms is a concerted effort among distinct bacterial clades with distinct glycan niches (e.g., (Xing et al., 2015)).

In this study, we provide evidence that the succession of bacterioplankton clades that we reported for the 2009 North Sea spring phytoplankton bloom re-occurred during the spring blooms from 2010 to 2012. We tested whether the bacterioplankton clades and their associated CAZyme repertoires differ from year to year or exhibit recurrent patterns. We analyzed spring bacterioplankton community composition via 16S rRNA catalyzed reporter deposition fluorescence *in situ* hybridization (CARD-FISH) and 16S rRNA gene tag sequencing, as well as gene function repertoires by deep metagenome sequencing. Our efforts have culminated into the as of yet highest resolved dataset capturing the response of planktonic bacteria to marine spring phytoplankton blooms and have allowed identification of recurring patterns that might ultimately lead to an explanatory model for bacterioplankton succession dynamics during spring algae blooms.

2.4 Materials and methods

2.4.1 Phytoplankton and physicochemical data

Physicochemical parameters (Supplementary file 1 - online) and phytoplankton data (Supplementary file 2 - online) were assessed in subsurface water on a weekday basis as part of the Helgoland Roads LTER time series. Details on the acquisition of these data

have been described previously (Teeling et al., 2012). The Helgoland Roads time series is accessible via the public database Pangaea (http://www.pangaea.de).

2.4.2 Bacterioplankton

Sampling of bacterioplankton was carried out as described previously (Teeling et al., 2012). In brief, surface seawater samples were taken at the long-term ecological research station 'Kabeltonne' (54° 11.3′ N, 7° 54′ E) at the North Sea island Helgoland using small research vessels (http://www.awi.de/en/expedition/ships/more-ships.html) and processed in the laboratory of the Biological Station Helgoland within less than two hours after sampling.

Biomass of free-living bacteria for DNA extraction was harvested on 0.2 µm pore sized filters after pre-filtration with 10 µm and 3 µm pore sized filters to remove large debris and particle-associated bacteria. By contrast, cells for microscopic visualization methods were first fixed by the addition of formaldehyde to sampled seawater, which was then filtered directly onto 0.2 µm pore sized filters. All filters were stored at -80°C until further use.

2.4.3 Microscopy: total cell counts, CARD-FISH

Assessment of absolute cell numbers and bacterioplankton community composition was carried out as described previously (Thiele et al., 2011). To obtain total cell numbers, DNA of formaldehyde fixed cells filtered on 0.2 µm pore sized filters was stained with 4',6-diamidino-2-phenylindole (DAPI). Fluorescently labeled cells were subsequently counted on filter sections using an epifluorescence microscope. Likewise, bacterioplankton community composition was assessed by catalyzed reporter deposition fluorescence *in situ* hybridization (CARD-FISH) of formaldehyde fixed cells on 0.2 µm pore sized filters. DAPI and CARD-FISH cell counts are summarized in Supplementary file 3 (online) and the corresponding probes in Supplementary file 4 (online).

2.4.4 16S rRNA V4 gene tag sequencing

Surface seawater samples were collected on bi-monthly to bi-weekly time scales from January 2010 to December 2012 at Helgoland roads. 500 ml of each sample were subjected to fractionating filtration as described above using 10, 3 and 0.2 µm pore size polycarbonate membrane filters (Millipore, Schwalbach, Germany). DNA of the 0.2–3 µm

fraction was extracted from filters as described previously (Sapp et al., 2007) and quantified using the Invitrogen (Carlsbad, CA, USA) Quant-iT PicoGreen dsDNA reagent as per manufacturer's instructions. Concentrations ranged from <1 to 20 µg DNA/ml.

50 µl aliquots of each sample were pipetted into 96-well plates and sent to the Department of Energy (DOE) Joint Genome Institute (JGI, Walnut Creek, CA, USA) for amplification and sequencing as follows: Sample prep was done on a PerkinElmer (Waltham, MA, USA) Sciclone NGS G3 Liquid Handling Workstation capable of processing 96 plate-based samples in parallel, utilizing the 5 PRIME (Gaithersburg, MD 20878, USA) HotMasterMix amplification kit and custom amplification primers targeting the V4 region of the 16S rRNA gene using 515F (5'-GTGCCAGCMGCCGCGGTAA-3') and 806R (5'-GGACTACHVGGGTWTCTAAT-3') (Caporaso et al., 2011). Primers also contained Illumina adapter sequences and a barcode index. PCR reactions were set up in 75 µl with 1x HotMasterMix (5 PRIME) with final concentrations of 0.4 µg/µl BSA and 0.2 µM of each primer. This volume was split into triplicate 25 µl reactions for independent amplification and then pooled to reduce PCR bias. Prepared amplicon libraries were normalized, multiplexed into a single pool per plate and quantified using the KAPA Biosystems (Wilmington, MA, USA) next-generation sequencing library qPCR kit on a Roche (San Francisco, CA, USA) LightCycler 480. Libraries were sequenced on an Illumina (San Diego, CA, USA) MiSeq sequencer using the Reagent Kit v3 and 2x250 bp chemistry. The resulting sequences are available from the DOE-JGI GOLD database (Reddy et al., 2015) as part of the COGITO project (Gp0056779) and from the NCBI short read archive (SRA) (SRA278189).

2.4.5 16S rRNA gene tag analysis

Roche 454 16S rRNA gene tags from 2009 MIMAS (Microbial Interactions in Marine Systems) project (Teeling et al., 2012) were reanalyzed for comparison with the Illumina-based COGITO extension project from subsequent years 2010–2012. The 2009 datasets was generated using the primers Bakt_314F (5' CCTACGGGNGGCWGCAG 3') and Bakt_805R (5' GACTACHVGGGTATCTAATCC 3') (Herlemann et al., 2011). The forward primers of both datasets target distinct regions, but the reverse primers target the same region. Hence, only those 454 reads sequenced from the 805 direction were reanalyzed for comparison. For 2010–2012, raw MiSeq paired-end reads (2x250 bp) were merged and filtered using illumina-utils (https://github.com/meren/illumina-utils) to retain only read pairs without mismatches in the overlapping regions. These high-quality Illumina tags and the 454 tags were then processed separately but with the same methods via the SILVAngs pipeline (Quast et al., 2013), which includes additional quality filtering steps via alignment as well as length, ambiguity and homopolymer filters. Sequences

were dereplicated at 100% identity and then globally clustered at 98%. Representative OTUs were classified to genus level against the SILVA (Quast et al., 2013) v119 database using BLAST with a similarity threshold = (sequence identity + alignment coverage) / 2 $\geq$ 93%. The SAR92 clade was reclassified according to SILVA v123. Reads were mapped against representative OTUs to obtain final abundance counts. For the purpose of this study, OTUs were collapsed based on shared taxonomy no higher than the genus level.

For MIMAS samples, we retained a total of 110,995 454 reads across 7 samples with an average of 16,000 per sample. After SILVAngs quality filtering, 110,866 remained for clustering. 6,102 representative OTUs were identified and 107,708 total sequences were assigned to a relative in the database during classification within the 93% similarity threshold. The final abundance matrix collapsed on shared taxonomic classification contained 500 unique taxa.

In total, 20,869,432 paired raw MiSeq reads were obtained across 142 samples from 2010–2012 COGITO samples. 15,016,350 merged reads with no mismatches in the overlapping region were retained with an average of 106,000 per sample. Reads were randomly sub-sampled to 40,000 tags per sample to reduce computational demands. In total 6,120,000 tags were submitted to the SILVAngs pipeline. After additional quality filtering, 6,116,021 sequences were clustered at 98% and the resulting 935,006 representative OTUs were classified. A total of 5,676,259 sequences were assigned to a relative in the database within the 93% similarity threshold. The final abundance matrix collapsed on shared taxonomy no higher than the genus level contained 1,995 unique taxa (Supplementary file 5 - online).

2.4.6 Metagenome sequencing

Total community DNA of 2009 samples (02/11/09; 03/31/09; 0407/09; 04/14/09; 06/16/09) was sequenced on the 454/Roche FLX Ti platform as described previously (Teeling et al., 2012). Metagenome sequencing of 2010–12 samples (03/03/2010; 04/08/10; 05/04/10; 05/18/10; 03/24/11; 04/28/11; 05/26/11; 03/08/12; 04/16/12; 05/10/12) was performed at the DOE Joint Genome Institute on the Illumina HiSeq2000 platform. Libraries were created from 100 ng environmental DNA per sample that was sheared to 270 bp using a Covaris E210 (Covaris, Woburn, MA, USA) and size selected using SPRI beads (Beckman Coulter, Indianapolis, IN, USA). The fragments were treated with end-repair, A-tailing, and ligation of Illumina compatible adapters (IDT, Coralville, IA, USA) using the KAPA-Illumina library creation kit. The libraries were quantified using KAPA Biosystem's next-generation sequencing library qPCR kit

and run on a Roche LightCycler 480 real-time PCR instrument. The quantified libraries were then prepared for sequencing on the Illumina HiSeq sequencing platform utilizing a TruSeq PE Cluster Kit v3, and Illumina's cBot instrument to generate a clustered flowcell for sequencing. Sequencing was performed on the Illumina HiSeq2000 sequencer using TruSeq SBS sequencing Kits, v3, following a 2x150 bp indexed run recipe.

Raw reads were screened against Illumina artifacts with kmer size of 28, step size of 1. Reads were subsequently trimmed from both ends using a minimum quality cut-off of 3; reads with three or more N's or with average quality score <Q20 were removed. In addition, reads <50 bp were removed. The remaining quality-filtered Illumina reads were assembled using SOAPdenovo v1.05 (Luo et al., 2012) at a range of kmers (81, 85, 89, 93, 97, 101) with default settings (options: -K 81 -p 32 -R -d 1). Contigs generated by each assembly (6 total contig sets), were de-replicated using JGI in house Perl scripts. Contigs were then sorted into two pools based on length. Contigs <1,800 bp were re-assembled using Newbler (Life Technologies, Carlsbad, CA, USA) in attempt to generate larger contigs (options: -tr, -rip, -mi 98, -ml 80). Contigs >1,800 bp as well as the contigs from the Newbler assembly were combined using minimus 2 (options: -D MINID=98 -D OVERLAP=80) from the AMOS package (http://sourceforge.net/projects/amos). Read depths were estimated based on read mapping with bbmap (http://bio-bwa.sourceforge.net). The metagenome study information is available from the DOE-JGI GOLD database (study: Gs0000079). The unassembled reads are available from the NCBI SRA (see Supplementary file 8 - online), and the assembled and annotated metagenome datasets from the IMG/M system (Markowitz et al., 2014).

2.4.7 Metagenome analysis

The DOE-JGI MAP v.4 annotation pipeline (Huntemann et al., 2015) was used for initial metagenome gene prediction and annotation. The annotated metagenomes were loaded in the IMG/M system as of mid 2014, and subsequently imported into a GenDB v2.2 annotation system (Meyer et al., 2003) for taxonomic classification and data mining.

All genes were searched against the NCBI non-redundant protein database (as of June 17th, 2014) using USEARCH v6.1.544 (Edgar, 2010), against the Pfam v25 database (Finn et al., 2014) using HMMER v3 (Punta et al., 2012), for signal peptides using SignalP v3.0 (Nielsen et al., 1999) and for transmembrane helices using TMHMM v2.0c (Krogh et al., 2001). CAZymes were automatically annotated based on HMMER searches against the Pfam v25 and dbCAN (Yin et al., 2012) databases and BLAST (Altschul et al., 1990) searches against the CAZy database (Cantarel et al.,

2009, Lombard et al., 2014) using E-value cut-offs that were specifically adjusted for each CAZyme family (Supplementary file 11 - online). Genes were only annotated as CAZymes when at least two of the search results were congruent, and CAZymes were only analyzed for contigs ≥500 bp.

Taxonomic classification of the metagenome sequences into taxonomically coherent bins ('taxobins') was carried out with a modified version of the Taxometer approach described in (Teeling et al., 2012). Taxometer consolidates predictions of a set of individual sequence classification tools into a consensus using a weighted assessment on seven selected ranks (superkingdom, phylum, class, order, family, genus, species) of the NCBI taxonomy (http://www.ncbi.nlm.nih.gov/Taxonomy). We combined taxonomic information inferred from (i) Pfam hits using the CARMA3 approach (Gerlach & Stoye, 2011), (ii) BLASTp hits using the KIRSTEN approach((Teeling et al., 2012); supp. data), and (iii) mapping of quality-filtered (illumina-utils; https://github.com/meren/illumina-utils/) Illumina reads to selected reference sequences. In contrast to the original Taxometer approach we omitted signature-based classification with Self-Organizing Maps and mapping of reads containing partial 16S rRNA gene sequences. The prediction tools that were used are outlined below:

We used the HMMER-based module of CARMA3 (not the BLAST-based module) that infers taxonomy of sequences by post-processing genes with HMMER3 hits to the Pfam database. The basic principle is to apply a reciprocal search technique to reduce the number of identified matches and thereby to improve taxonomic classification quality.

KIRSTEN (Kinship Relationship Reestablishment) infers taxonomy of sequences by post-processing BLASTp hits to the NCBI nr database by means of rank-based evaluations on all levels of the NCBI taxonomy with an increasing stringency from the superkingdom down to the species level. On each taxonomic level, all occurring taxa are weighted by the sum of their BLASTp bit scores. When the taxon with the highest weight exceeds an adjustable threshold, the process continues towards the next taxonomic level. The threshold increases with each taxonomic level, i.e., the algorithm becomes more critical while it progresses. For this study, we substituted BLASTp by the UBLAST module of USEARCH 6.1 (Edgar, 2010) with an E-value cutoff of E-10 and maximum hit count of 500.

SMALT (http://www.sanger.ac.uk/resources/software/smalt) was used to map metagenome reads on a manually compiled set of 49 reference genomes and a streamlined version of the NCBI nr database. Both, the reference genomes and the sequences selected from the NCBI nr database were selected based on habitat-specific information. This was done manually for the reference genomes and automatically for the NCBI nr

database as follows: Each of the hits from the UBLAST search during KIRSTEN analysis can be associated with multiple taxa, since in nr redundant sequences from different taxa are merged. We used this information to extract all taxa associated with a given hit, then combined the taxa of all hits and finally extracted all sequences of these taxa from Genbank. This way a sample-specific streamlined subset of Genbank was generated that greatly sped up the mapping process. Only metagenome reads with at least 95% identity to any sequence in the sample-specific sequence database were used. Reads were subsequently back-mapped to contigs. Since contigs consist of many reads, this way, contigs were associated with multiple taxonomic paths. Taxonomic paths resulting from classification with reference genomes and the habitat-specific streamlined nr were concatenated. Paths representing less than 1% of the reads of a given contig were discarded.

Finally, Taxometer was used to combine all gene-based predictions from CARMA3 and KIRSTEN and the mapping results for each contig, and to infer a consensus taxonomy. Taxonomic predictions were possible for 94.6% of the contigs above 1 kbp. The results are summarized in Supplementary file 9 (online).

2.4.8 Statistical analyses

The Spearman rank correlation test (Supplementary file 6 - online) was used to test for correlations between *Bacteroidetes* clade abundances and environmental variables (chlorophyll *a*, temperature, salinity, silicate, phosphate, nitrate, nitrite, ammonia) and phytoplankton abundances (classes: diatoms, dinoflagellates, coccolithophorids, silicoflagellates, flagellates, ciliates, green algae; species: *Mediophyxis helysia*, *Thalassiosira nordenskioeldii*, *Chaetoceros debilis* and *C. minimus*, *Rhizosolenia styliformis*, *Chattonella* and *Phaeocystis*). *Bacteroidetes* and phytoplankton numbers were transformed to log-scale for better comparison. Linear regressions (Supplementary file 7 - online) were done using stepwise forward regression model by using the log transformed *Bacteroidetes* and phytoplankton abundances. All statistical analyses were performed using the software Sigma-Plot 12 (SYSTAT, Santa Clara, CA, USA).

2.5 Results

2.5.1 Sampling site characteristics

The samples were taken at Helgoland Island about 40 km offshore in the southeastern North Sea in the German Bight at the station 'Kabeltonne' (54° 11.3′ N, 7° 54′ E;

(Fig. 2.1) between the main island and the minor island, Düne (German for 'dune'). Water depths at this site fluctuate from 6 to 10 m over the tidal cycle. During most of the year, a westerly current transports water from the English Channel alongside the Dutch and Frisian coast to Helgoland, but water around the island is also influenced by nutrient inputs from the rivers Weser and Elbe and from the northern North Sea (Wiltshire et al., 2010). During the 2009 to 2012 study period, the lowest water temperatures were measured in mid to late February (min. 2010: 1.1°C; max. 2009: 3.4°C), followed by a continuous increase until a peak in August (min. 2011: 18.0°C; max. 2009: 18.7°C) (Supplementary file 1 - online).

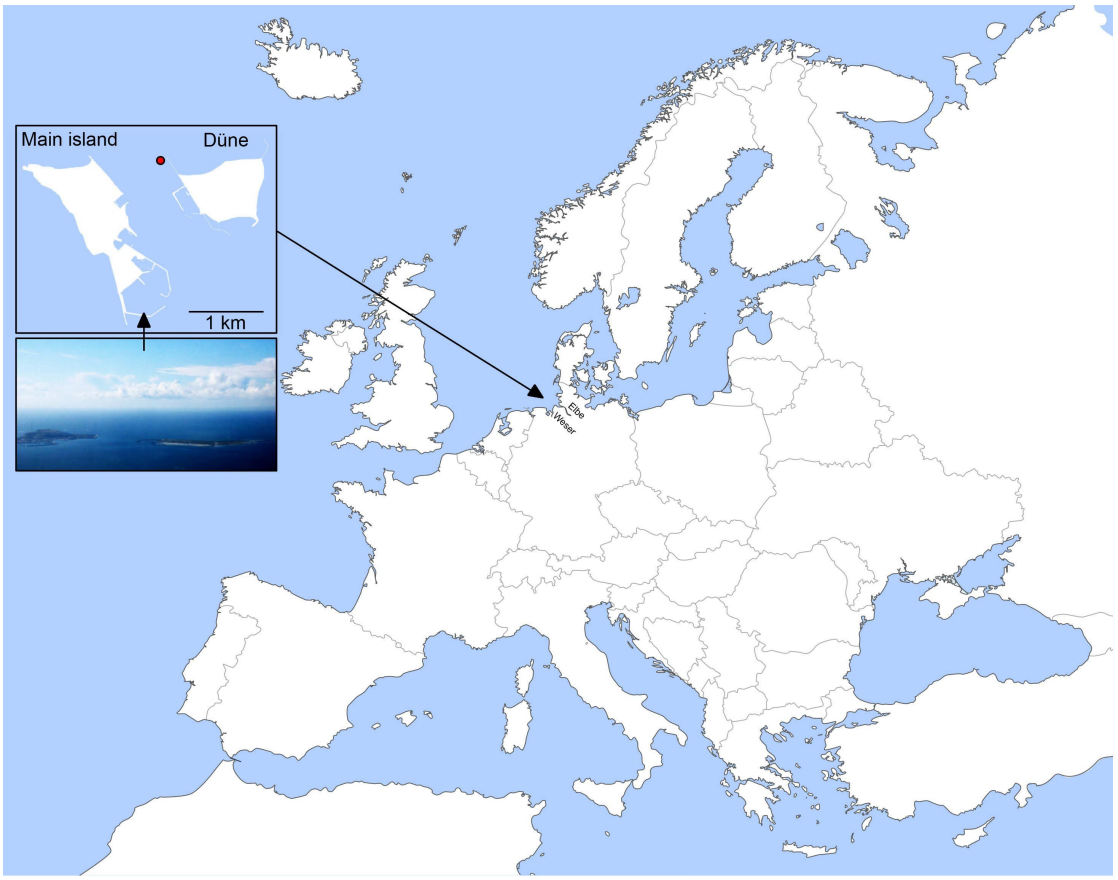


Figure 2.1: Location of Helgoland Island (ca. 40 km offshore the northern German coastline) and the long-term ecological research site 'Kabeltonne' (red circle: 54° 11.3′ N, 7° 54′ E) in the German Bight of the North Sea.

2.5.2 Phytoplankton - diversity and bloom characteristics

Spring phytoplankton blooms in the North Sea typically develop during March and reach highest intensities during April and May. The highest chlorophyll *a* concentrations are usually observed at the coastlines including the area around Helgoland Island (Fig. 2.2). North Sea spring blooms are thus large-scale phenomena that are, however,

influenced by local conditions, such as riverine inputs. At Helgoland island, spring phytoplankton blooms started around mid March when water temperatures surpassed 3 to 5°C (Fig. 2.3A–H; Supplementary file 1 - online). The diatoms *Chaetoceros debilis* and *Chaetoceros minimus*, *Mediopyxis helysia*, *Rhizosolenia styliformis* and *Thalassiosira nordenskioeldii*, the silicoflagellate *Chattonella*, the haptophyte *Phaeocystis* and dinoflagellates dominated these blooms in terms of cell numbers (Fig. 2.3I–L; Supplementary file 2 - online). Relative abundances of these algae varied in no apparent order during the observed blooms, and we have yet to understand the factors that determine these variations. The sizes of the dominant algae taxa are different, with *Chaetoceros minimus* and in particular *Phaeocystis* spp. featuring the smallest cells and *Mediopyxis helysia* and *Rhizosolenia styliformis* featuring the largest cells. Spherical *Phaeocystis* spp. cells for example have estimated biovolumes of ∼50 to 250 μm^3, whereas elipsoid cylindrical *Mediopyxis helysia* cells have a biovolume of ∼82,000 μm^3 and for *Rhizosolenia styliformis* even a biovolume of ∼282,000 μm^3 has been reported (Loebl et al., 2013, Olenina et al., 2006). Considering biomass, the blooms were largely dominated by the diatoms *T. nordenskioeldii* and *M. helysia* and the silicoflagellate *Chattonella*. Blooms of these three algae were bimodal in all years with dominance of first *T. nordenskioeldii* followed by *Chattonella* in 2009 (Fig. 2.3I), *T. nordenskioeldii* followed by *M. helysia* in 2010 (Fig. 2.3J), *M. helysia* followed by *Chattonella* in 2011 (Fig. 2.3K) and a pronounced bimodal bloom of *Chattonella* species in 2012 (Fig. 2.3L). All blooms were accompanied by a notable decrease of silicate (Fig. 2.3A–D; Supplementary file 1 - online), which diatoms use for frustule formation (see Yool & Tyrrell (2003) for the controlling effect of silicate on diatom abundance).

Bloom maximum intensities decreased from 2009 to 2012 with chlorophyll *a* maxima (measured by fluorescence) reaching 28 mg/m^3 (day 82), 18 mg/m^3 (day 125), 15 mg/m^3 (day 116), and 11 mg/m^3 (day 114) in each respective year (Fig. 2.3A–D; Supplementary file 1 - online). Maximum bacterioplankton cell counts were observed either close to or after the chlorophyll *a* maxima (Supplementary file 3 - online). In 2009, the bacterioplankton peaked at $3.5x10^6$ cells/ml (day 118), 36 days after the Chl *a* maximum. In 2010, the peak in Chl *a* was broader and only four days ahead of the bacterioplankton peak abundance of $2.4x10^6$ cells/ml (day 129). In 2011, the Chl *a* peak was only two days ahead of the bacterioplankton peak abundance of $2.3x10^6$ cells/ml (day 118), whereas it was nine days ahead in 2012, where the bacterioplankton peaked at $3.2x10^6$ cells/ml (day 123).

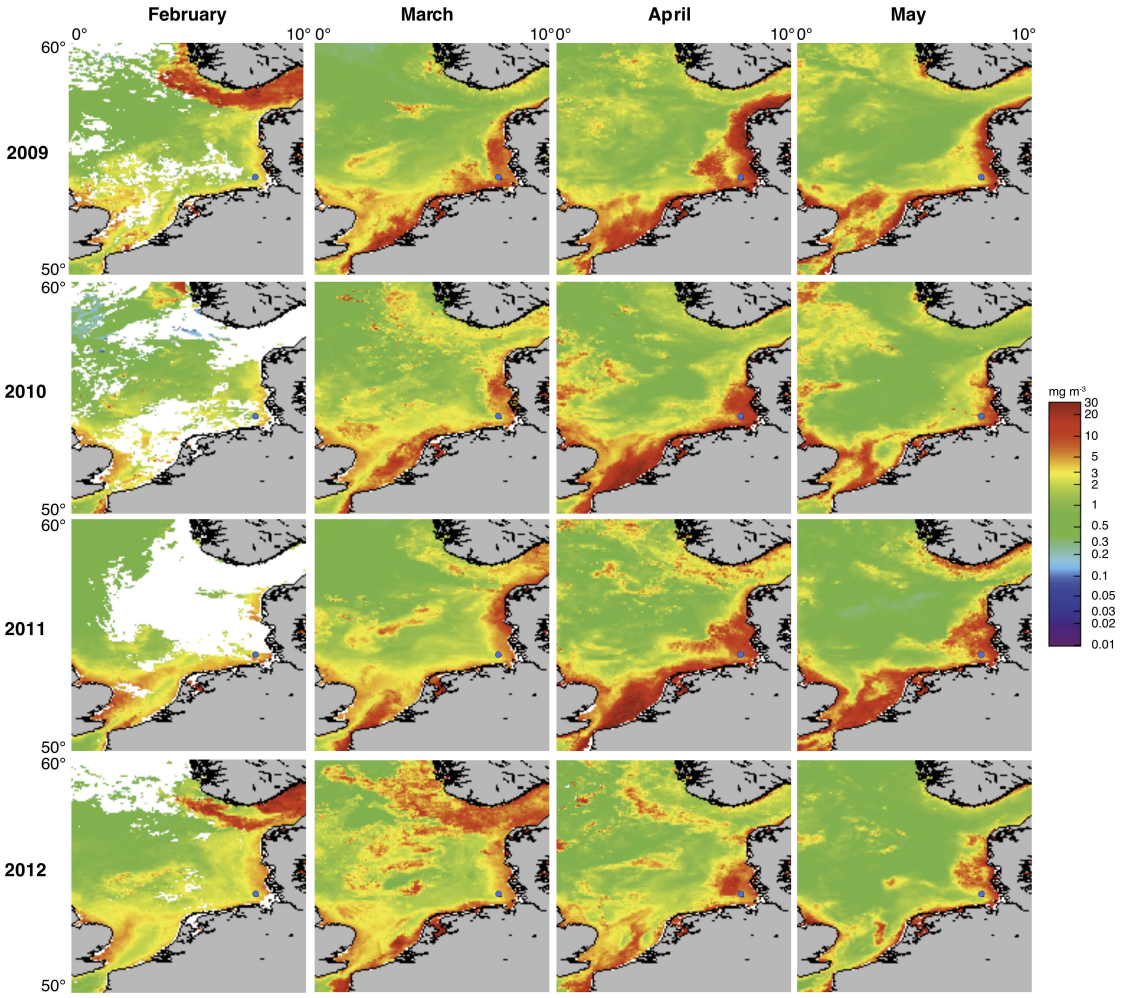


Figure 2.2: Satellite chlorophyll *a* measurements. Data are shown for the southern North Sea for the months February to May (monthly averages) of the years 2009 to 2012. Images were retrieved from the GlobColour website using the 'extended Europe area' at full resolution (1 km) as merged products of weighted averages from the following sensors: MERIS, MODIS AQUA, SeaWIFS and VIIRS. See GlobColour website for details (http://hermes.acri.fr). The position of Helgoland Island is indicated by a blue dot.

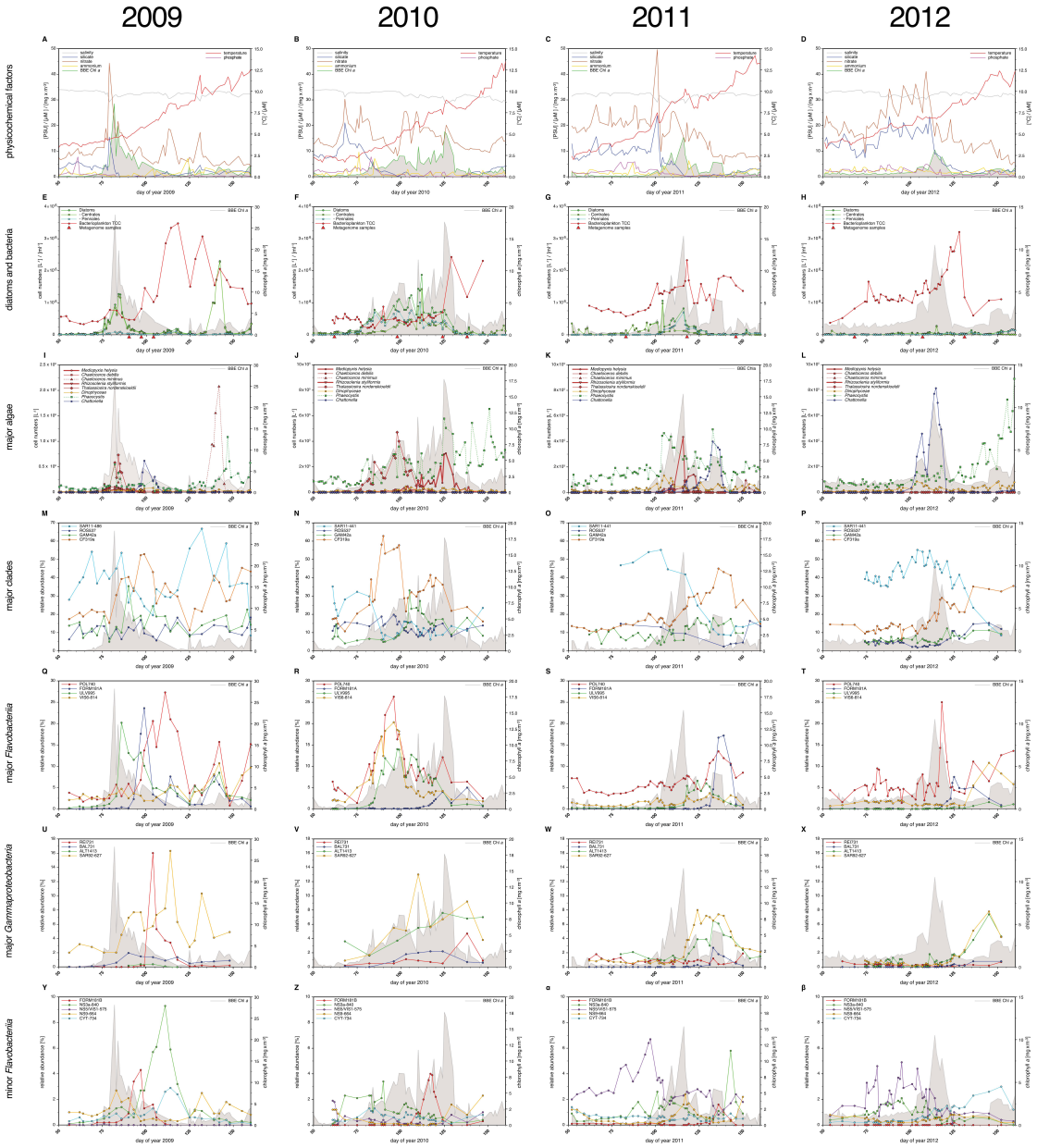


Figure 2.3: **Physicochemical parameters, phytoplankton composition and bacterioplankton composition as assessed by CARD-FISH.** A-D: Physicochemical measurements. Left ordinate: salinity (PSU), silicate (µM), nitrate (µM), ammonium (µM) and chlorophyll *a* (mg/m^3); right ordinate: temperature (°C) and phosphate (µM). E-H: Counts of the diatom groups and total bacterioplankton cell counts (TCC). I-L: Microscopic cell counts of the most abundant algae genera (red: diatoms; orange: dinoflagellates: green: haptophytes; blue: silicoflagellates). Algae with large cells and thus large biovolumes are depicted by bold solid lines and algae with small cells are represented by dotted lines. M-ß: Recurrent bacterioplankton clades as assessed by CARD-FISH (catalyzed reporter deposition-fluorescence *in situ* hybridization). M-P: Major bacterial groups: SAR11-486 and SAR11-441: alphaproteobacterial SAR11-clade; ROS537: alphaproteobacterial *Roseobacter* clade; GAM42a: *Gammaproteobacteria*; CF319a: *Bacteroidetes*. Q-T: Major *Flavobacteriia* clades: POL740: genus *Polaribacter*; FORM181A: genus *Formosa*; ULV995: genus *Ulvibacter*; VIS6-814: genus-level clade VIS6 within the family *Cryomorphaceae-Owenweeksia*; U-X: Major *Gammaproteobacteria* clades: REI731: genus *Reinekea*; BAL731: genus*Balneatrix*; ALT1413: families *Alteromonadaceae* and *Colwelliaceae*; SAR92-627: genus-level clade SAR92. Y-ß: Minor *Bacteroidetes* clades: FORM181B: species-specific for *Formosa* sp. Hel1_33_131; NS3a-840: NS3 marine group; NS5/VIS1-575: VIS1 genus-level clade within the NS5 marine group; NS9-664: NS9 marine group; CYT-734: *Cytophagia* clade *Marinoscillum*. See https://doi.org/10.7554/eLife.11888.005 for electronic version and complete figure legend.

2.5.3 Bacterioplankton - diversity and bloom characteristics

DAPI and CARD-FISH cell staining (Supplementary file 3 - online, Supplementary file 4 - online) showed that SAR11 dominated the bacterioplankton community in winter, but with the onset of each spring bloom relative abundances of *Bacteroidetes* followed by *Gammaproteobacteria* increased and finally surpassed those of the SAR11 (Fig. 2.3M–P). *Bacteroidetes* reached higher maximum relative abundances than *Gammaproteobacteria*, 40% (2012) to 60% (2010) as compared to 10% (2012) to 30% (2010), respectively.

Bacteroidetes genera *Polaribacter*, *Formosa*, and VIS6, a genus-level clade within the family *Cryomorphaceae* (Gómez-Pereira et al., 2010), peaked each year with relative abundances well above 5%, reaching relative abundances of up to 25% sometimes within less than a week (Fig. 2.3Q–T). *Ulvibacter* reached similar peak abundances with the exception of 2012, where this genus never surpassed 2% and ranged below 1% most of the time. Within *Gammaproteobacteria* the genus-level SAR92 clade responded notably in all years increasing from background levels below 1% to relative abundance of 8% to 16%. Members of the *Alteromonadales* families *Alteromonadaceae* and *Colwelliaceae* bloomed in three (2010: 8%; 2011: 6%; 2012: 7%), and genus *Reinekea* in two (2009: 16%; 2010: 5%) of the years (Fig. 2.3U–X). Some less abundant, but nevertheless recurrent taxa included the genus *Balneatrix* within *Gammaproteobacteria* with relative abundances up to 2% (Fig. 2.3U–X). Minor recurring groups of *Bacteroidetes* (Fig. 2.3Y-ß) included the NS3a marine group (9% in 2009 and 6% in 2011), the genus-level VIS1 clade within the NS5 marine group detected before the Chl *a* peaks of 2011 (7%) and 2012 (5%), and the *Cytophagia* clade *Marinoscillum* that reached 1–3% abundance most years after initial blooms.

We used complementary 16S rRNA gene tag sequencing for the detection of bacterioplankton clades that were not recovered by CARD-FISH probes (Supplementary file 5 - online). Relative proportions of 16S tags from distinct clades correlated for the most part those inferred from CARD-FISH cell counts (Fig. 2.4A-P), but members of SAR11 were substantially underreported - a known limitation of the 806R primer used in V4 amplification for 2010 to 2012 samples (Apprill et al., 2015). Additional abundant clades detected in the 16S amplicon data comprised the *Flavobacteriia* genus *Tenacibaculum* (Fig. 2.4U–X) that bloomed in 2010 (read frequencies of max. ∼12%) and 2011 (max. ∼5%). Within *Gammaproteobacteria*, clades with read frequencies $\geq 5\%$ in at least one year comprised the genera *Aeromonas*, *Glaciecola*, *Pseudoalteromonas*, *Pseudomonas*, *Psychrobacter* and the SAR86 and ZD0405 clades (Fig. 2.4Q–T). Within the alphaproteobacterial *Rhodobacteriaceae*, high abundances of '*Candidatus* Planktomarina temperata' (DC5-80-3 lineage) and the NAC11-7 clade were detected reaching 6–21% and 7–19% of the tag data, respectively (Fig. 2.4U–X). Also within *Alphaproteobacteria*

the genus *Sulfitobacter* peaked with a read frequency of ∼7% in 2010 (Fig. 2.4V), and within *Betaproteobacteria* the order *Methylophilales* (dominated by OM43 clade members) was detected with high relative abundances of up to ∼10% before blooms, which decreased with bloom progression. *Verrucomicrobia* (Fig. 2.4U–X) were detected with decreasing peak read frequencies of 7.7% (2010), 5.0% (2011) and 2.9% (2012). This decrease corresponds to decreasing bloom intensities, which supports a proposed role of *Verrucomicrobia* in polysaccharide decomposition (e.g., (Martinez-Garcia et al., 2012)).

Within *Bacteroidetes*, *Gammaproteobacteria* and *Rhodobacterales* a total of eleven clades peaked during at least two of the four spring blooms with relative cell abundances or, for those clades that were not assessed by CARD-FISH, relative read frequencies $\geq 5\%$. These were six *Flavobacteriia* clades (*Formosa*, *Polaribacter*, NS3a marine group, *Tenacibaculum*, *Ulvibacter*, VIS6 clade *Cryomorphaceae*), three *Gammaproteobacteria* clades (*Alteromonadaceae/Colwelliaceae*, *Reinekea* and SAR92), and two *Roseobacter* clades (DC5-80-3 and NAC11-7).

Each year a succession was observed within the *Flavobacteriia* and *Gammaproteobacteria* clades. The succession in the *Flavobacteriia* was more pronounced than in the *Gammaproteobacteria*, but the sequence of clades varied. Spearman rank correlation analyses revealed that the abundances of the most prominent *Flavobacteriia* clades were for the most part correlated with multiple algae groups and physiochemical factors (Supplementary file 6 - online). According to linear regression analyses, the strongest abiotic predictors were temperature, salinity, silicate and nitrate, and the strongest biotic predictors were *Phaeocystis* spp. haptophytes, *Rhizosolenia* spp., *Chaetoceros debilis*, and *Chaetoceros minimus* diatoms and the silicoflagellate *Chattonella* (Supplementary file 7 - online). It should be noted though that linear regressions were computed based on log-transformed abundance data and not algae volumes (which were not measured). Thus, the influence of the rather small cell-sized algae such as *Chaetoceros minimus* is likely overestimated. Such limitations notwithstanding it is noteworthy that in no case a simple one-to-one relationship between specific algae and specific bacterioplankton groups was detected. The strongest significant ($p<0.05$) correlations were obtained for the *Ulvibacter* clade that was positively correlated with diatoms and haptophytes and negatively correlated with silicoflagellates. Further results comprised an opposite trend for the VIS1 clade of the NS5 marine group, and a correlation of *Polaribacter* and *Chattonella* abundances (see Supplementary file 6 - online for details).

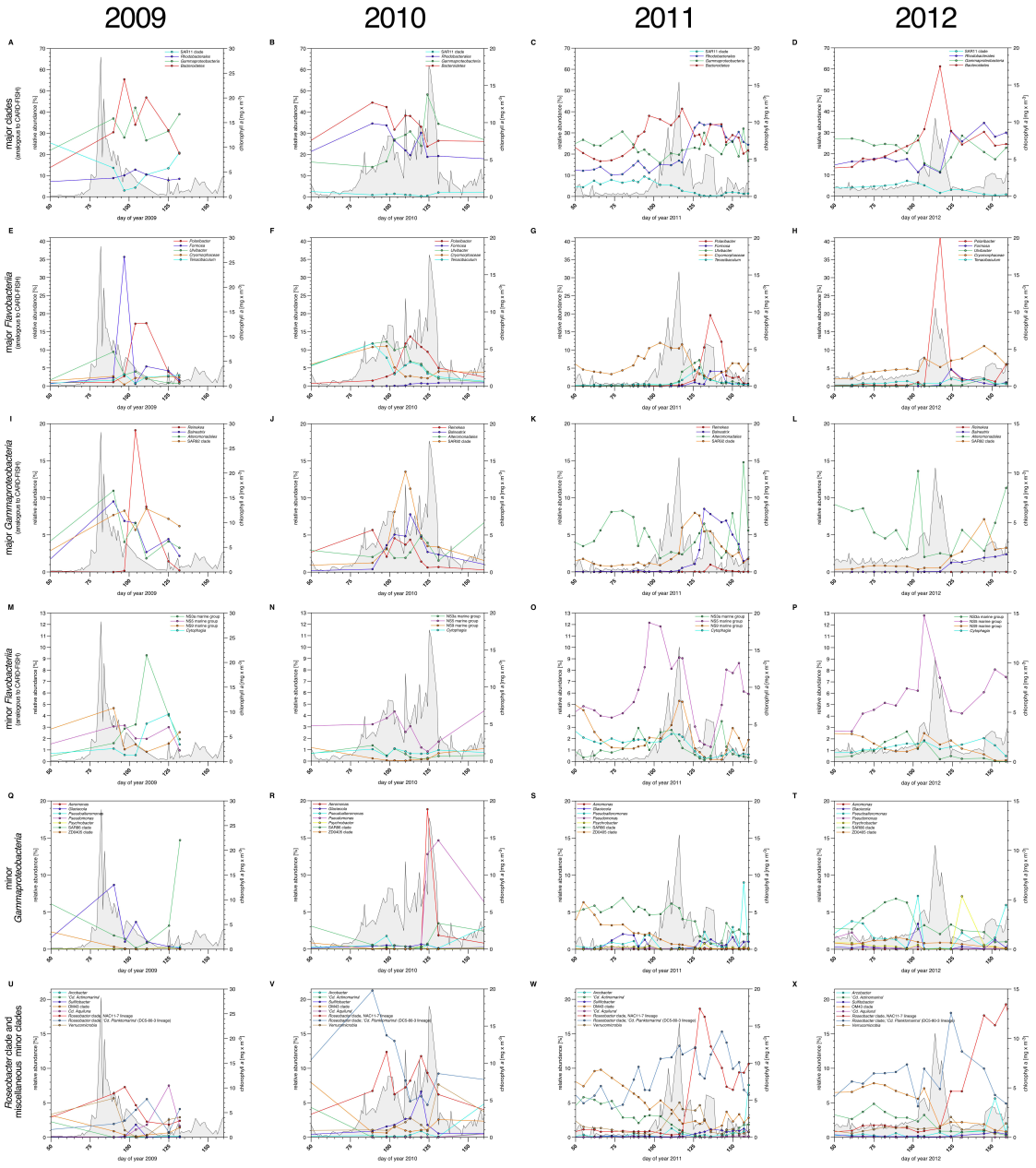


Figure 2.4: Bacterioplankton diversity as assessed by 16S rRNA gene tag sequencing. Time points from days 50 to 160 were plotted for all years. Panel A-P depict data that are analogous to the CARD-FISH data presented in Fig. 2.3, with addition of the *Flavobacteriia* genus *Tenacibaculum* (E-H). Panels Q-X show minor *Gammaproteobacteria* clades (Q-T) and *Roseobacter* clades together with miscellaneous other minor clades (U-X) that were not tested by CARD-FISH probes. See https://doi.org/10.7554/eLife.11888.006 for electronic version and complete figure legend.

2.5.4 Bacterioplankton - genetic repertoires

In total, 16 metagenomes of free-living bacterioplankton (0.2 – 3 μm) were generated from time points before, during and after spring phytoplankton blooms, six during 2009 using the 454 FLX Ti platform that were published previously (Teeling et al., 2012)

and ten during 2010–2012 using the Illumina HiSeq2000 platform. Most of the 454 (0.5–4 pico titer plates / metagenome) and all of the Illumina metagenomes (1 lane / metagenome; 2x150 bp) were deeply sequenced (Supplementary file 8 - online) with final assembled contigs of up to 96 kbp and 458 kbp, respectively.

Taxonomic classification of the metagenome contigs resulted in identification of major bloom-associated clade sequence bins (Supplementary file 9 - online), including *Formosa*, *Polaribacter*, the NS3a and NS5 marine groups, and *Cryomorphaceae* of the *Flavobacteriia* and *Alteromonadales*, *Reinekea*, *Glaciecola* and the SAR92 clade of the *Gammaproteobacteria*. Classification was poor, however, for *Ulvibacter* (*Flavobacteriia*) and *Balneatrix* (*Gammaproteobacteria*), most likely since the only available reference genomes (unidentified eubacterium SCB49; *Balneatrix alpica*) were too distant from North Sea representatives. Clone libraries from 2009 (Teeling et al., 2012) indicated 16S rRNA similarities of only 94% and 91%, respectively, for these two clades. Other abundant clades comprised the betaproteobacterial *Burkholderiales* and *Methylophilales* (including the OM43 clade), the alphaproteobacterial SAR116 and *Roseobacter* NAC11-7 clade, and the gammaproteobacterial SAR86 and ZD0405 clades. Lower abundant clades comprised, amongst others, the OM60 (NOR5) group, the AEGEAN-169 group, and *Sulfitobacter*.

We plotted contig GC content versus coverage to evaluate our taxonomic classification, which in some cases allowed us to assess the coherence of some of the clades (Fig. 2.5). For example, *Reinekea* (Fig. 2.5D,H,I) and the NS5 marine group (Fig. 2.5H,L,M,O) were mostly represented by distinct clusters, whereas *Polaribacter* (Fig. 2.5D,H,I) was almost always represented by at least two clusters indicating the presence of sub-populations. In general, the number of clusters increased from pre-bloom to mid-bloom situations and decreased slightly towards late bloom situations and notably towards post-bloom situations. This tendency was more evident in 2009, the year with the highest bloom intensity and the largest number of metagenome samples spanning a broader timespan (Fig. 2.5A–F). It is noteworthy that high *in situ* abundance did not always correlate with good metagenome assemblies. SAR11 for example, while highly abundant in all metagenome datasets, yielded few large contigs, possibly due to population heterogeneity and presence of hyper-variable regions described in sequenced SAR11 genomes (Wilhelm et al., 2007).

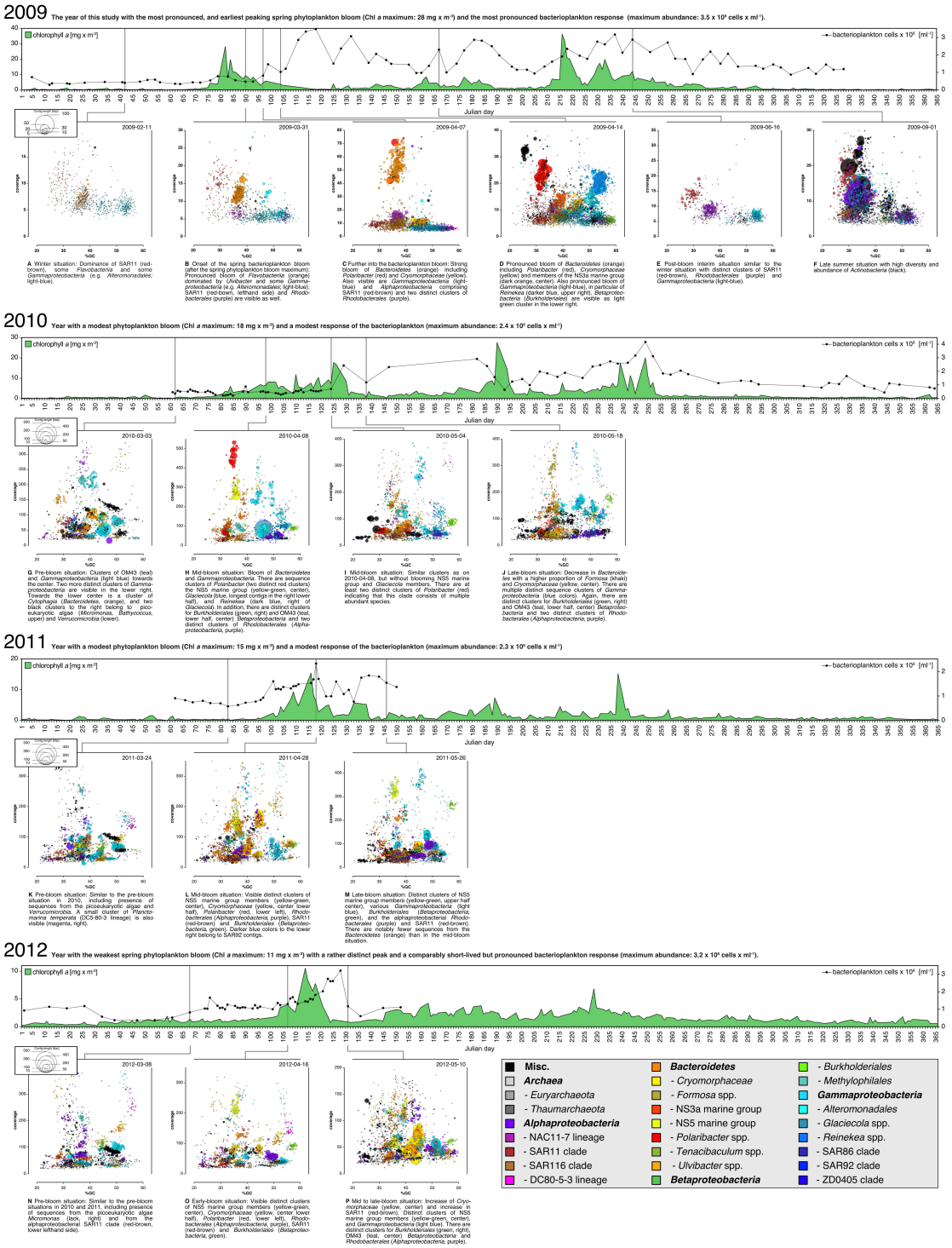


Figure 2.5: Taxonomic classification of bacterioplankton metagenomes. Metagenome contigs are visualized as bubbles with radii that are proportional to their lengths and colors that indicate their predicted taxomomic affiliations. These bubbles are drawn in planes that are defined by the contig's GC content and coverage values. Colors are restricted to selected abundant taxa (see legend below) to highlight distinct clusters, mostly from the *Bacteroidetes*, *Alphaproteobacteria*, *Betaproteobacteria* and *Gammaproteobacteria*. Likewise only contigs are shown that exceed a minimum length of 2,750 bp for pyrosequencing data (2009) and 15,000 bp for llumina data (2010-2012), respectively. Sparse contigs with very high coverage or GC content below 20% or above 60% were also excluded from visualizations. The 16 metagenomes are shown arranged in order on yearly timescales that depict chlorophyll *a* contents as proxies for phytoplankton abundance. See https://doi.org/10.7554/eLife.11888.007 for electronic version and complete figure legend.

Analyses of functional genes identified in the bacterioplankton metagenomes revealed that increases in *Flavobacteriia* relative abundance during blooms was always accompanied by an increase in community-wide CAZyme gene frequency as well as an increase in the diversity of CAZyme families (Fig. 2.6A–C). As blooms subsided, CAZyme frequencies also declined. This was most pronounced for more densely sampled years in 2009 and 2010. In 2012 the decline in CAZymes was not captured as the last metagenome sample was taken before the bloom decline (Fig. 2.6A, x-axis).

The 20 glycoside hydrolase families with the highest mean abundances during bloom dates were, in descending order, GH3, 23, 13, 16, 103, 73, 92, 2, 17, 30, 5, 20, 36, 65, 29, 1, 42, 31, 81, and 18 (Fig. 2.6D). While most of these families comprise a diverse range of functions, they are indicative for the hydrolysis of certain glucans, xylans, mannans (GH92), cellulose (GH5), fucans (GH29), and peptidoglycan (GH23, 103, 73, and 18). The abundant families GH3, 16, 17 and 5 comprise β-1,3-glucanases and the family GH30 β-1,6-glucanases. Both enzymes are involved in the cleavage of chrysolaminarin. Chrysolaminarin, a mostly linear β-1,3-glucan with occasional β-1,6 branches, is the storage polysaccharide in diatoms and thus one of the most abundant polysaccharides on Earth.

The ten families of carbohydrate-binding modules with the highest mean abundances during blooms were, in descending order, CBM50, 32, 6, 9, 48, 11, 22, 57, 20, and 4 (Fig. 2.6E). These CBM domains are indicative for the binding of peptidoglycan (CBM50), galactans (CBM32), glycogen (CBM48), β-1,3- and β-1,4-glucans (CBM6, 11, 4), xylan (CBM6, 9, 22, 4), and cellulose (CBM6, 4). This suggests a pronounced specialization of the bacterial community in the acquisition of storage polysaccharides (β-1,3-glucans such as chrysolaminarin; α-1,4-glucans such as starch and glycogen) and cell wall polysaccharides (xylose, cellulose and peptidoglycan) of both algae and bacteria. By contrast, CBMs for the binding of algal TEP, which consists predominantly of fucose- and rhamnose-rich anionic sulfated heteropolysaccharides (Passow, 2002), were only found in lower abundances, e.g., CBM47 (fucose) and CBM67 (rhamnose).

Among carbohydrate esterase families, CE11, 4, 1, and 14 exhibited the highest mean abundances during blooms (Fig. 2.6F). The only known function for family CE11 is that of the UDP-3-O-acyl-N-acetylglucosamine deacetylase, an enzyme involved in lipid A biosynthesis. The second and third most abundant CE families 4 and 1 comprise a number of enzymatic functions including deacetylation of peptidoglycan, xylan, and chitin.

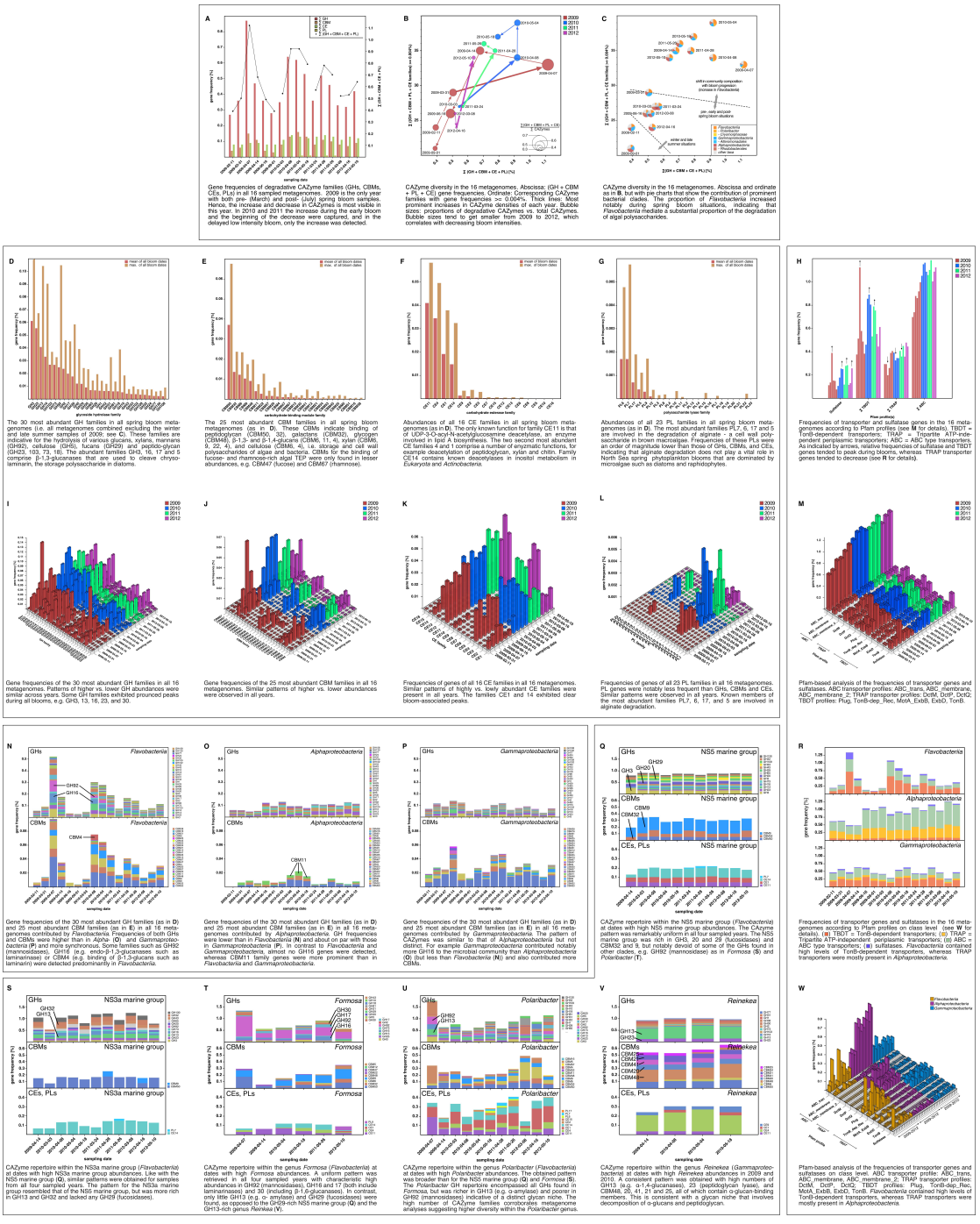


Figure 2.6: Metagenome functional analyses: CAZyme, sulfatase and transporter gene frequencies. CAZymes were predicted as consensus of searches against the CAZy, dbCAN and Pfam databases with custom E-value cutoffs (Supplementary file 11 - online). Sulfatase and transporter genes were predicted based on HMMER searches against the Pfam databases with an E-value cutoff of E-5. Gene frequencies were computed as [(sum of average coverage of target genes) ∗100 / (sum of average coverage of all genes)]. All dates in the graphs are in the format [yyyy-mm-dd]. See https://doi.org/10.7554/eLife.11888.008 for electronic version and complete figure legend.

Finally, the polysaccharide lyase families PL6, 7, 17 and 5 constituted the most abundant PL families during bloom dates (Fig. 2.6G). Known members of these PL families are

all involved in the usage of alginate. However, gene frequencies of these PL families were an order of magnitude below those of abundant GHs, CBMs, and CEs, indicating that alginate degradation does not play a vital role in North Sea spring phytoplankton blooms. Alginate is a cell wall constituent in brown macroalgae, however, the microalgae that dominate North Sea spring blooms are devoid of alginate.

For all of these GH, CBM, PL, and CE families, we observed remarkably similar gene frequency patterns during all four spring blooms, often with peaks in the same families (Fig. 2.6I–L). Alongside CAZymes, sulfatase and TonB-dependent transporter (TBDT) gene frequencies also increased during blooms while tripartite ATP-independent periplasmic (TRAP) transporter genes showed an almost opposite trend (Fig. 2.6H,M). A class-level analysis revealed that sulfatases and TBDT genes were predominantly present in *Bacteroidetes*, whereas TRAP transporters were mostly present in *Alphaproteobacteria* (Fig. 2.6R,W). The observed shifts in sulfatase, TBDT and TRAP transporter frequencies hence reflect the shifts in relative abundance between these two classes. This is in agreement with our study on the 2009 spring bloom (Teeling et al., 2012) and furthermore demonstrates recurrence of this phenomenon during four consecutive years.

Class-level analyses of the most abundant GH and CBM families (Fig. 2.6N–P) showed that *Flavobacteriia* not only contributed more total CAZymes to the microbial community than *Alpha-* and *Gammaproteobacteria*, but also exhibited a tighter coupling between GH and CBM genes with highly similar abundance profiles (Fig. 2.6N). The distribution of families was also more uneven in *Flavobacteriia*, indicative of a more pronounced substrate specialization compared to *Alpha-* and *Gammaproteobacteria*. GH92 (mannosidase) and GH16 (including β-1,3-glucanase) genes, for example, were predominantly found in *Flavobacteriia*, possibly indicating more readiness to decompose mannans and chrysolaminarin.

Metagenome taxonomic classification provided sufficient data for the analysis of CAZyme repertoires of the flavobacterial NS5 marine group, NS3a marine group, *Formosa*, *Polaribacter*, and *Cryomorphaceae*, and the gammaproteobacterial *Alteromonadales* and *Reinekea* clades (Supplementary file 10 - online). For most of these clades, the analyses revealed fingerprint-like patterns, which corroborates the hypothesis that these clades have distinct glycan niches that are relatively stable across years. For example, the NS5 marine group (Fig. 2.6Q) was rich in GH3, 20 and 29 (fucosidases), but notably devoid of GH92 (mannosidases). By contrast, Formosa and *Polaribacter* clades (Fig. 2.6T,U) contained higher abundances of GH92 genes. The *Formosa* CAZyme profile was also characterized by high proportions of GH16, 17, and 30 families, which all contain enzymes that can decompose chrysolaminarin. *Polaribacter* contained a broader set of CAZymes that included all families found in *Formosa*, however, *Polaribacter* was richer in GH13

(e.g., α-amylase) and poorer in GH92 (mannosidases) than *Formosa*. Likewise, the GH repertoires of the NS3a and NS5 marine groups were similar (Fig. 2.6Q,S), but the NS3a marine group was richer in GH13 and 32 and devoid of GH29 family fucosidases. The high number of CAZyme families in *Polaribacter* corroborated metagenome bin analyses that suggested a higher diversity within this clade. CAZyme gene frequencies were much lower in the *Cryomorphaceae* than in the other investigated *Flavobacteriia* clades with GH frequencies barely exceeding 0.5% (Fig. 2.6 – figure supplement 1-online). This suggests a different ecophysiological niche and a distinct role of the *Cryomorphaceae* during phytoplankton blooms.

For *Gammaproteobacteria*, recurring patterns were detected for the prominent *Alteromonadales* and *Reinekea* clades. *Alteromonadales* contained some of the GH families that play important roles during phytoplankton blooms, such as GH13 and 16, but were notably poor in or even devoid of others, such as GH29 and GH92, respectively (Fig. 2.6 – figure supplement 2-online). In contrast to other prominent clades, we did not obtain sufficient metagenome sequences for *Reinekea* for all four years, but only for 2009 and 2010 (Fig. 2.6V). However, the *Reinekea* CAZyme patterns of 2009 and 2010 were well conserved with high proportions of GH23 and 13, and CBM48, 20, 41, 21, and 25. The GH23 family comprises peptidoglycan lyases and the GH13 family contains α-1,4-glucanases (e.g., α-amylase). CBM48, 20, 41, 21, and 25 all bind α-1,4-glucans such as starch and glycogen, and the ubiquitous CBM50 contains peptidoglycan-binding members. These results are consistent with a glycan niche that involves decomposition of external α-1,4-glucans and possibly peptidoglycan.

2.6 Discussion

Nutrient-poor marine surface waters are dominated by clades such as SAR11 and *Prochlorococcus*. Both feature small, reduced genomes and can use sunlight and small organic molecules (e.g., Gómez-Pereira et al. (2013)). The otherwise heterotrophic SAR11 use proteorhodopsin for supplemental phototrophy (Giovannoni et al., 2005) and phototrophic *Prochlorococcus* cyanobacteria are capable of supplemental uptake of amino acids (Zubkov et al., 2003) and glucose (Muñoz-Marín et al., 2013). Microbial communities in nutrient-rich 'green' surface oceans by contrast feature higher proportions of heterotrophic species that feed on more complex organic substrates. In particular during phytoplankton blooms, the release of algae-derived organic matter selects for fast growing species with genomic adaptations towards algal biomass remineralization. These are typically members of the *Flavobacteriia* and *Gammaproteobacteria* classes and the alphaproteobacterial *Roseobacter* clade. Similarly adapted species from these clades

compete for substrates during phytoplankton blooms with variation in which species prevail. Despite this stochastic effect, the most well-adapted species will be successful more often and thus exhibit patterns of annual recurrence.

During spring phytoplankton blooms at Helgoland in the North Sea, we observed recurrent bloom-associated abundance peaks of in particular flavobacterial clades, namely *Formosa*, *Polaribacter*, the NS3a marine group *Tenacibaculum*, *Ulvibacter*, and the *Cryomorphaceae* VIS6 clade. Within *Gammaproteobacteria Alteromonadaceae/Colwelliaceae*, *Reinekea*, and the SAR92 clade were clearly bloom-associated and recurrent as was *Methylophilales* within *Betaproteobacteria*, and '*Candidatus* Planktomarina temperata' from the DC5-80-3 lineage (a.k.a *Roseobacter* clade affiliated = RCA group) and the NAC11-7 lineage within the *Roseobacter* clade. It has already been shown that the abundant North Sea isolate '*Candidatus* Planktomarina temperata' RCA23T is associated with decaying phytoplankton (Giebel et al., 2011) and high abundances and in particular high activity have been reported during a spring phytoplankton bloom event in the North Sea of 2010 (Voget et al., 2015, Wemheuer et al., 2015). High activities of members of the RCA and the SAR92 clades during North Sea spring phytoplankton blooms have also been reported in 2009 (Klindworth et al., 2014) and 2010 (Wemheuer et al., 2015), just as an increase of *Bacteroidetes* of the genera *Marinoscillum* and *Polaribacter* during 2010 (Wemheuer et al., 2015).

Many of the other clades we report here (including some low abundance groups) have been found during blooms of dinoflagellates, including AEGEAN-169, *Alteromonadales*, NS3a marine group, NS5 marine group, OM43, OM60 (NOR5), SAR116, SAR86, and ZD0405 (Yang et al., 2015) or *Cryomorphaceae*, *Glaciecola* and *Sulfitobacter* (Tan et al., 2015). The OM43 clade (order *Methylophilales*) comprises methylotrophs known to feed on algae C1 compounds (Halsey et al., 2012), and it has been reported that *Sulfitobacter* species SA11 shares a mutually beneficial exchange of compounds with the diatom species *Pseudo-nitzschia multiseries* (Amin et al., 2015).

There is of course unaddressed diversity in all these clades. Some of the genera might be dominated by a single species while others might be more diverse with considerable variation in competitive success between bloom events. Nevertheless, the high level of recurrence in particular of flavobacterial clades indicates a strong selection of few clades highly adapted for the manipulation and uptake of specific and complex polysaccharides (and likely other biopolymers) and disagrees with substantial levels of sloppy-feeding by these bacteria that would allow other less adapted clades to arbitrarily reach high abundances via cross-feeding. This might be attributed to the capability of *Bacteroidetes* for very efficient macromolecule uptake as it has been recently shown for uptake of α-mannan by the human gut bacterium *Bacteroides thetaiotaomicron* (Cuskin et al.,

2015). This bacterium binds α-mannan macromolecules to its surface, followed by rapid cleavage into larger oligomers that are immediately imported via a TonB-dependent transporter (TBDT) into the periplasm without detectable loss. The bulk of the degradation into smaller molecules takes place in the periplasm where the substrate is secure from outside competitors before transport into the cytoplasm.

TonB-dependent transporters are not specific to *Bacteroidetes*, but it seems that only *Bacteroidetes* have evolved a functional coupling of SusC-like TBDT porins with SusD-like TonB-dependent receptors (TBDRs) that bind and guide the substrate to the porin. At least so far, only *Bacteroidetes* genomes feature characteristic *susCD* gene tandems. Within *Bacteroidetes* genomes such tandems are frequently found in so-called polysaccharide utilization loci (PULs; (Sonnenburg et al., 2010)). PULs are operons or regulons where one or more *susCD* gene tandems are co-located with CAZymes. Further frequent accessory genes in PULs encompass transcriptional regulators, proteases, transporter components and sulfatases. The latter are required for the desulfation of sulfated polysaccharides, which marine algae produce in large quantities. The diversity of PULs in marine *Bacteroidetes* genomes is high and largely unexplored, and so far only few PULs have been linked to dedicated algal polysaccharides (e.g., (Hehemann et al., 2012b, Kabisch et al., 2014, Xing et al., 2015)). The large, complex and efficient PUL uptake systems might explain why *Flavobacteriia* consistently outcompeted *Gammaproteobacteria* during the onset of all blooms.

It is noteworthy that we observed a shift in bacterioplankton biodiversity alongside a shift in functional gene repertoires for the major clades in all four years of this study. The abundance of CAZymes and sulfatases increased from pre- to mid-bloom situations and leveled off post-bloom. Likewise, similar abundance patterns were observed for the most abundant CAZyme families in all studied years.

We did observe an increase in the abundance of TBDRs during bloom situations, and we have shown previously that TBDRs are among the most abundantly expressed proteins during the bacterial mineralization of algae biomass (Teeling et al., 2012). The relevance of TBDRs in nutrient-rich oceanic regions has been also supported by *in situ* metaproteome studies of samples from the South Atlantic Ocean (in particular at coastal upwelling zones; (Morris et al., 2010)) and from the Antarctic Southern Ocean (Williams et al., 2013).

The recurrent patterns in bacterioplankton diversity and functional repertoires during the four studied spring blooms are remarkable in view of the variation among algae taxa. For example, even though algal biomass of the 2012 spring bloom was dominated by silicoflatellate *Chattonella* spp. and not by diatoms as was true for the three preceding years, the respective bacterioplankton communities were strikingly similar.

Likewise only few bacterioplankton taxa seemed to be weakly correlated with dedicated phytoplankton taxa, and no clear correlation was found between distinct bacterioplankton taxa and individual distinct diatom clades in statistical analyses (Supplementary file 6 - online and Supplementary file 7 - online). It seems that phytoplankton community composition did not exert a strong effect on the composition of the free-living non-phycosphere bacterioplankton community. Instead, the dominating algae (in terms of biomass) seemed to produce similar or perhaps identical types of substrates for specifically adapted clades of heterotrophic bacterioplankton. It is therefore conceivable that recurrence is more pronounced on the functional level than on the taxonomic level, since species from different taxa with similar ecophysiological niches might functionally substitute each other in different years. This hypothesis is supported by the bacterioplankton communities' similar CAZyme gene repertoires in the 2009 to 2012 spring blooms and in particular the consistency on class level that was almost unaffected by distinct blooming clades, yet needs to be further tested by deep metatranscriptome sequencing and metaproteomics during multiple spring phytoplankton blooms in future studies. Considering the extent of recurrence, our combined metagenome data (>5 million predicted proteins) should provide sufficient search space for such an analysis.

The existence of recurrent key players during North Sea spring phytoplankton blooms suggests that the bacterioplankton community composition during and after such blooms is governed by deterministic effects. We have shown before that temperature exerts a strong effect on North Sea bacterioplankton as it selects for temperature-dependent guilds, for example when comparing spring and summer blooms (Lucas et al., 2015). Within short-lived spring blooms, however, the supply of algae-derived organic matter is among the main factors that shape the bacterioplankton composition. In particular the different types of structurally distinct polysaccharides that algae produce in large quantities seem to exert such substrate-induced forcing. Since *Flavobacteriia* are more specialized on polysaccharides than *Gammaproteobacteria*, this would also explain, why *Flavobacteriia* dominated the recurrent clades (Fig. 2.3Q–T, Fig. 2.4E–J) and *Gammaproteobacteria* clades exhibited more stochastic peaks (Fig. 2.4Q–T).

At the beginning of a bloom, most available polysaccharides will be exopolysaccharides, but as the bloom commences and algae become senescent, more and more cellular algal substrates are released, culminating in the bloom's final die off phase. Bacteria will naturally consume the more degradable substrates such as storage polysaccharides (e.g., chrysolaminarin) first, and more recalcitrant substrates (e.g., branched and sulfated polysaccharides) later. TEP for example seems to undergo such selective feeding, as it has been suggested that in particular fucose-rich TEP is less readily degraded than

mannose and galactose rich TEP (for review see Passow (2002)). Such selective feeding creates an additional change in substrate availability and leads to a succession of substrate niches for specifically adapted bacterioplankton clades to grow.

2.6.1 Concluding remarks

Bacterioplankton communities during spring phytoplankton blooms in the coastal North Sea undergo swift and dynamic composition changes and thus are difficult to investigate. Nonetheless, we found clades that recurrently reached high abundances among *Flavobacteriia* (*Formosa*, *Polaribacter*, NS3a marine group, *Ulvibacter*, VIS6 clade *Cryomorphaceae*, *Tenacibaculum*), *Gammaproteobacteria* (*Alteromonadaceae/Colwelliaceae*, SAR92, *Reinekea*) and *Roseobacter* clade *Alphaproteobacteria* (DC5-80-3, NAC11-7). Recurrence was not only detectable on the taxonomic but also on the functional level with a highly predictable increase in TonB-dependent polysaccharide uptake systems and distinct CAZyme patterns. The niches of abundant bacterioplankton clades are more complex and manifold than the glycan niches that we explore in this study. CAZymes, however, have the advantage that they allow linking of gene repertoires and possible environmental functions in a way currently not feasible for other macromolecules such as proteins and lipids. Our results suggest that besides stochastic also deterministic effects influence phytoplankton-bacterioplankton coupling during blooms. They indicate that during spring phytoplankton blooms similar principles of resource partitioning and specialization are at play as within human gut microbiota that decompose fiber-rich plant material, albeit at a much larger scale. Rather than one-to-one interactions of particular phytoplankton and bacterioplankton species, the availability of substrates commonly occurring in microalgae caused the succession of free-living bacterioplankton clades.

2.7 Funding

This study was funded by the Max Planck Society. The work conducted by the US. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231.

2.8 Acknowledgements

The expert technical assistance of Y.-L.Chen and D. Berkelmann is acknowledged. This study was funded by the Max Planck Society. The work conducted by the US. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, is supported under Contract No. DE-AC02-05CH11231.

Chapter 3

Polysaccharide utilization loci of North Sea *Flavobacteriia* as basis for using SusC/D-protein expression for predicting major phytoplankton glycans

Lennart Kappelmann[1], Karen Krüger[1], Jan-Hendrik Hehemann[1,2], Jens Harder[1], Stephanie Markert[3,4], Frank Unfried[3,4], Dörte Becher[5], Nicole Shapiro[6], Thomas Schweder[3,4], Rudolf I. Amann[1], Hanno Teeling[1]

[1]Max Planck Institute for Marine Microbiology, Bremen, Germany
[2]Zentrum für Marine Umweltwissenschaften, Bremen, Germany
[3]Pharmaceutical Biotechnology, University Greifswald, Greifswald, Germany
[4]Institute of Marine Biotechnology, Greifswald, Germany
[5]Institute for Microbiology, University Greifswald, Greifswald, Germany
[6]DOE Joint Genome Institute, Walnut Creek, CA, USA

Published in *The ISME Journal* (DOI: 10.1038/s41396-018-0242-6).

Contributions:
Experimental concept and design: 15%
Acquisition of experimental data: 10%
Data analysis and interpretation: 15%
Preparation of figures and tables: 5%
Drafting of the manuscript: 10%

Electronic figure versions and supplementary material are available at https://doi.org/10.1038/s41396-018-0242-6.

3.1 Abstract

Marine algae convert a substantial fraction of fixed carbon dioxide into various polysaccharides. *Flavobacteriia* that are specialized on algal polysaccharide degradation feature genomic clusters termed polysaccharide utilization loci (PULs). As knowledge on extant PUL diversity is sparse, we sequenced the genomes of 53 North Sea *Flavobacteriia* and obtained 400 PULs. Bioinformatic PUL annotations suggest usage of a large array of polysaccharides, including laminarin, α-glucans, and alginate as well as mannose-, fucose-, and xylose-rich substrates. Many of the PULs exhibit new genetic architectures and suggest substrates rarely described for marine environments. The isolates' PUL repertoires often differed considerably within genera, corroborating ecological niche-associated glycan partitioning. Polysaccharide uptake in *Flavobacteriia* is mediated by SusCD-like transporter complexes. Respective protein trees revealed clustering according to polysaccharide specificities predicted by PUL annotations. Using the trees, we analyzed expression of SusC/D homologs in multiyear phytoplankton bloom-associated metaproteomes and found indications for profound changes in microbial utilization of laminarin, α-glucans, β-mannan, and sulfated xylan. We hence suggest the suitability of SusC/D-like transporter protein expression within heterotrophic bacteria as a proxy for the temporal utilization of discrete polysaccharides.

3.2 Introduction

Half of global net primary production is oceanic and carried out mostly by small, unicellular phytoplankton such as diatoms (Field et al., 1998). Polysaccharides account for up to 50% of algal biomass (Kraan, 2012) and can be found as intracellular energy storage compounds, as structural components of their cell walls (Kloareg & Quatrano, 1988), or as secreted extracellular transparent exopolymeric substances (Hoagland et al., 1993). They can be composed of different cyclic sugar monomers linked by either α- or β-glycosidic bonds at different positions and can be substituted by different moieties (e.g., sulfate, methyl, or acetyl groups), making them the most structurally diverse macromolecules on Earth (Laine, 1994).

Many members of the bacterial phylum *Bacteroidetes*, including marine representatives of the class *Flavobacteriia*, are specialized on polysaccharide degradation. They feature distinct polysaccharide utilization loci (PULs, Bjursell et al. (2006)), i.e., operons or regulons that encode the protein machinery for binding, degradation and uptake of a type or class of polysaccharides. Polysaccharides are initially bound by outer membrane proteins and cleaved by endo-active enzymes into oligosaccharides suitable for transport through

the outer membrane. Oligosaccharides are bound at the interface of SusCD complexes. SusD-like proteins are extracellular lipoproteins and SusC-like proteins constitute integral membrane beta-barrels termed TonB-dependent transporters (TBDTs). Glenwright et al. (2017) showed that these two proteins form a 'pedal bin' complex in *Bacteroides thetaiotaomicron*, with SusD acting as a lid on top of the SusC-like TBDT. Upon binding of a ligand, the SusD lid closes and conformational changes lead to substrate release into the periplasm. Here, further saccharification to sugar monomers takes place that are taken up into the cytoplasm via dedicated transporters.

Besides the characteristic *susCD*-like gene pair, *Bacteroidetes* PULs contain various substrate-specific carbohydrate-active enzymes (CAZymes), such as glycoside hydrolases (GHs), polysaccharide lyases (PLs), carbohydrate esterases (CEs), carbohydrate-binding modules (CBMs), and proteins with auxiliary functions. PULs of human gut *Bacteroidetes* and their capacity to degrade various land plant polysaccharides have been thoroughly investigated (e.g., ref. Martens et al. (2011)), but knowledge on marine polysaccharide degradation is sparse. Many polysaccharides in marine algae differ from those in land plants. Green macroalgae contain ulvans, red macroalgae contain agars, carrageenans and porphyrans, brown algae contain alginates, fucans and laminarin, and diatom microalgae contain chrysolaminarin and sulfated mannans, all of which are presumably absent in land plants (Popper et al., 2011). Likewise, many algae feature anionic, sulfated polysaccharides that require sulfatases for degradation.

A systematic inventory of the structural diversity of algal polysaccharides has not yet been achieved. We do not have a good understanding of the associated diversity of PULs in marine *Bacteroidetes*. Also only few PULs have so far been linked to their polysaccharide substrate. Examples include an agar/porphyran-specific PUL (Hehemann et al., 2012b) that human gut *Bacteroidetes* acquired from marine counterparts (Hehemann et al., 2010), an alginate-specific PUL in *Zobellia galactanivorans* DsiJ^{T} (Thomas et al., 2012), alginate- and laminarin-specific PULs in *Gramella forsetii* KT0803 (Kabisch et al., 2014), a similar laminarin-specific PUL in *Polaribacter* sp. Hel1_33_49 (Xing et al., 2015), and a complex carrageenan degradation regulon in *Z. galactanivorans* DsiJ^{T} (Ficko-Blean et al., 2017). Few overarching comparative genomic studies exist (Barbeyron et al., 2016b, Xing et al., 2015), focusing largely on overall CAZyme repertoires.

Pioneering studies on structural elucidation of polysaccharides from microalgae were performed (Ford & Percival, 1965, Rees & Welsh, 1977), but precise microalgal polysaccharide structures remain mostly unresolved (for review, see ref. Hoagland et al. (1993)), because they require sophisticated methods (Le Costaouëc et al., 2017). PUL analysis of heterotrophic bacteria co-occurring with phytoplankton could serve as an alternative

starting point to advance insight into the structures of marine polysaccharides and to understand their microbial decomposition.

Here we present a comparative analysis of PULs from 53 newly sequenced *Flavobacteriia* isolated from the German Bight, comprising a total of 400 manually determined PULs. Based on these data we investigated whether SusC- and SusD-like sequences can be linked to distinct predicted polysaccharides. Using environmental metaproteome data we show how SusC/D homolog expression may be used to assess the presence of marine polysaccharides during North Sea spring blooms.

3.3 Material and methods

3.3.1 Isolation and sequencing of North Sea *Flavobacteriia*

Flavobacteriia were sampled at the North Sea Islands Helgoland and Sylt as described previously (Hahnke & Harder (2013) and Hahnke et al. (2015), Supplementary Table S1 - online). Also included were the previously sequenced *Gramella forsetii* KT0803 (Bauer et al., 2006), *Polaribacter* spp. Hel1_33_49 and Hel1_85 (Xing et al., 2015), and the *Formosa* spp. Hel1_33_131 and Hel3_A1_48. The remaining 48 genomes were sequenced at the Department of Energy Joint Genome Institute (DOE-JGI, Walnut Creek, CA, USA) in the framework of the Community Sequencing Project No. 998 COGITO (Coastal Microbe Genomic and Taxonomic Observatory). Forty genomes were sequenced using the PacBio RSII platform exclusively, whereas eight isolates were sequenced using a combination of Illumina HiSeq 2000/2500 and PacBio RSII. All these genomes are GOLD certified at level 3 (improved high-quality draft) and are publicly available at the DOE-JGI Genomes OnLine Database (GOLD, Reddy et al. (2015)) under the Study ID Gs0000079.

3.3.2 Gene and PUL annotation

Initial annotations of the genomes of *Polaribacter* spp. Hel1_33_49 and Hel1_85 and *Formosa* spp. Hel1_33_131 and Hel3_A1_48 were performed using the RAST annotation system (Aziz et al., 2008). All other genomes were annotated using the DOE-JGI Microbial Annotation Pipeline (MGAP, (Huntemann et al., 2015)). These annotations were subsequently imported into a GenDB v2.2 annotation system (Meyer et al., 2003) for refinement and additional annotations based on similarity searches against multiple databases as described previously (Mann et al., 2013).

SusC- and SusD-like proteins were annotated by the DOE-JGI MGAP, which uses the TIGRfam model TIGR04056 to detect SusC-like proteins and the Pfam models 12741, 12771, and 14322 to detect SusD-like proteins. CAZymes were annotated based on HMMer searches against the Pfam v25 (Finn et al., 2014) and dbCAN 3.0 (Yin et al., 2012) databases and BLASTp searches (Altschul et al., 1990) against the CAZy database (Lombard et al., 2014). CAZymes were annotated only as such when at least two of the database searches were positive based on family-specific cutoff criteria that were described previously (Teeling et al., 2016). Selected sulfatases were annotated using the SulfAtlas database v1.0 (Barbeyron et al., 2016a). Peptidases were annotated using BLASTp searches against the MEROPS 9.13 database (Rawlings et al., 2012) using the default settings of $E \leq 10^{-4}$.

PULs were manually detected based on the presence of CAZyme clusters, which in most cases also featured co-occurring *susCD*-like gene pairs as previously suggested (Bjursell et al., 2006). In some cases, the sequence similarity of a TBDT was too low to be considered SusC-like, no SusD homolog was present or the entire *susCD*-like gene tandem was missing. These operons were still counted as PULs and are regarded as incomplete subtypes (Hemsworth et al., 2016).

3.3.3 Gene expression analyses of *Flavobacteriia*-rich North Sea bacterioplankton using metaproteomics

During spring phytoplankton blooms of 2009 to 2012, 14 surface seawater biomass samples were collected at the long-term ecological research station 'Kabeltonne' (54° 11.3′ N, 7° 54.0′ E) off the German North Sea island Helgoland as previously described in detail (Teeling et al., 2012, 2016). Biomass was collected on 0.2 µm pore sized filters after pre-filtration with 10 and 3 µm pore sized filters. Metagenome sequencing was done using the 454 FLX Ti platform for 2009 and the Illumina HiSeq 2000 platform for 2010 to 2012 samples (Teeling et al., 2016).

Corresponding metaproteome analyses were performed from biomass obtained from the same water samples. Protein extraction from 0.2 µm filtered bacterioplankton biomass and separation was carried out as described previously (Teeling et al., 2012) with the modification that gel lanes were cut into 10 equal pieces prior to tryptic digestion (1 µg/ml, Promega, Madison WI, USA) and subsequent mass spectrometric detection in an LTQ Orbitrap Velos mass spectrometer (Thermo Fisher, Bremen, Germany). The mass spectrometry proteomics data have been deposited to the ProteomeXchange Consortium via the PRIDE partner repository (Vizcaíno et al., 2016); data set identifiers: PXD008238, 10.6019/PXD008238.

Mass spectrometric data were analyzed using Sequest v27r11 (Thermo Fisher Scientific, San Jose, CA, USA). Searches were carried out against a forward-decoy database of all proteins from all metagenome samples combined. This non-redundant database was constructed from all predicted protein-coding genes of all metagenomes (6,194,278 sequences) using the uclust option of USEARCH v6.1.544 (Edgar, 2010); options: cluster_fast; nucleotide identity 0.99; maxhits 5; maxrejects 30) and contained 3,212,324 sequences. Common laboratory contaminants were included in all databases. Technical duplicates of each sample were searched together (including all 20 subsamples) to obtain averaged spectral counts. Validation of protein- and peptide identifications was performed with Scaffold v4 (Proteome Software Inc, Portland, OR, USA) using the parameters previously described (Teeling et al., 2012), and normalized spectral abundance factors (%NSAF) were calculated (Zhang et al., 2010) to allow for semi-quantitative analyses (Supplementary Table S2 - online). The NSAF quantitation measure is commonly used in non-gel-based label-free shotgun proteomics. In brief, a %NSAF of 1 corresponds to 1% of all mass-adjusted spectral count data in a given proteomic experiment.

3.3.4 SusC/D homolog tree reconstruction

We constructed trees from SusC- and SusD-like protein sequences of the isolates' PULs (SusC: 369; SusD: 361). Sequences were aligned using MAFFT v7.017 (Katoh & Standley, 2013) using the G-INS-i algorithm and BLOSUM62 matrix with default gap open penalty (1.53) and offset (0.123) values. Maximum-likelihood trees were constructed using FastTree 2.1.5 (Price et al., 2010) with default settings.

3.4 Results

3.4.1 High genomic and phylogenetic diversity in isolated marine *Flavobacteriia*

The 53 flavobacterial isolates cover a broad range of the *Flavobacteriia* class within the phylogenetic tree based on full-length 16S rRNA genes (Fig. 3.1). The strains fall into several clusters that can be linked to characteristic genomic features (Supplementary Table S1 - online). Genome sizes ranged from 2.02 Mbp (*Formosa* sp. Hel3_A1_48) to 5.98 Mbp (*Aquimarina* sp. MAR_2010_214), with an average of 3.83 Mbp. One of the clusters was dominated by isolates obtained from the retentates of seawater filtered through 20 µm particle nets (8 out of 12; Fig. 3.1; Supplementary Table S1 - online). These species feature mostly larger genomes (average 4.5 Mbp) and are likely associated with microalgae. Forty-seven of the 53 strains have two to four 16S rRNA operons,

with the notable exception of the three *Tenacibaculum* strains possessing six (strains MAR_2009_124 and MAR_2010_205) and seven (strain MAR_2010_89), respectively.

The capacity of the isolates to degrade polysaccharides varied widely as indicated by the number of degradative CAZymes per Mbp and predicted PULs per genome. On average, we identified 7.5 PULs per genome and 55 degradative CAZymes (Supplementary Table S1 - online). Strains of the putative microalgae-associated cluster differed with on average 83.3 degradative CAZymes, almost twice as many PULs per genome (14.2) and many sulfatase genes, indicating an extended capacity for the degradation of sulfated polysaccharides (average of 28.2 sulfatases, with a maximum of 95 sulfatases in *Zobellia amurskyensis* MAR_2009_138). The other strains had an average of 46.8 degradative CAZymes and 5.5 PULs. Eleven isolates possessed less than three PULs, contained few (≤ 3) or no sulfatases and were exclusively isolated from surface seawater or pore water. They likely target rather simple, non-sulfated polysaccharides and peptides. This strategy is emphasized by their high peptidase:CAZyme ratio of 1.81, compared with an average ratio of 0.95 for isolates with > 10 PULs. Still it is noteworthy that numbers of PULs and degradative CAZymes varied considerably, even within isolates of the same genus.

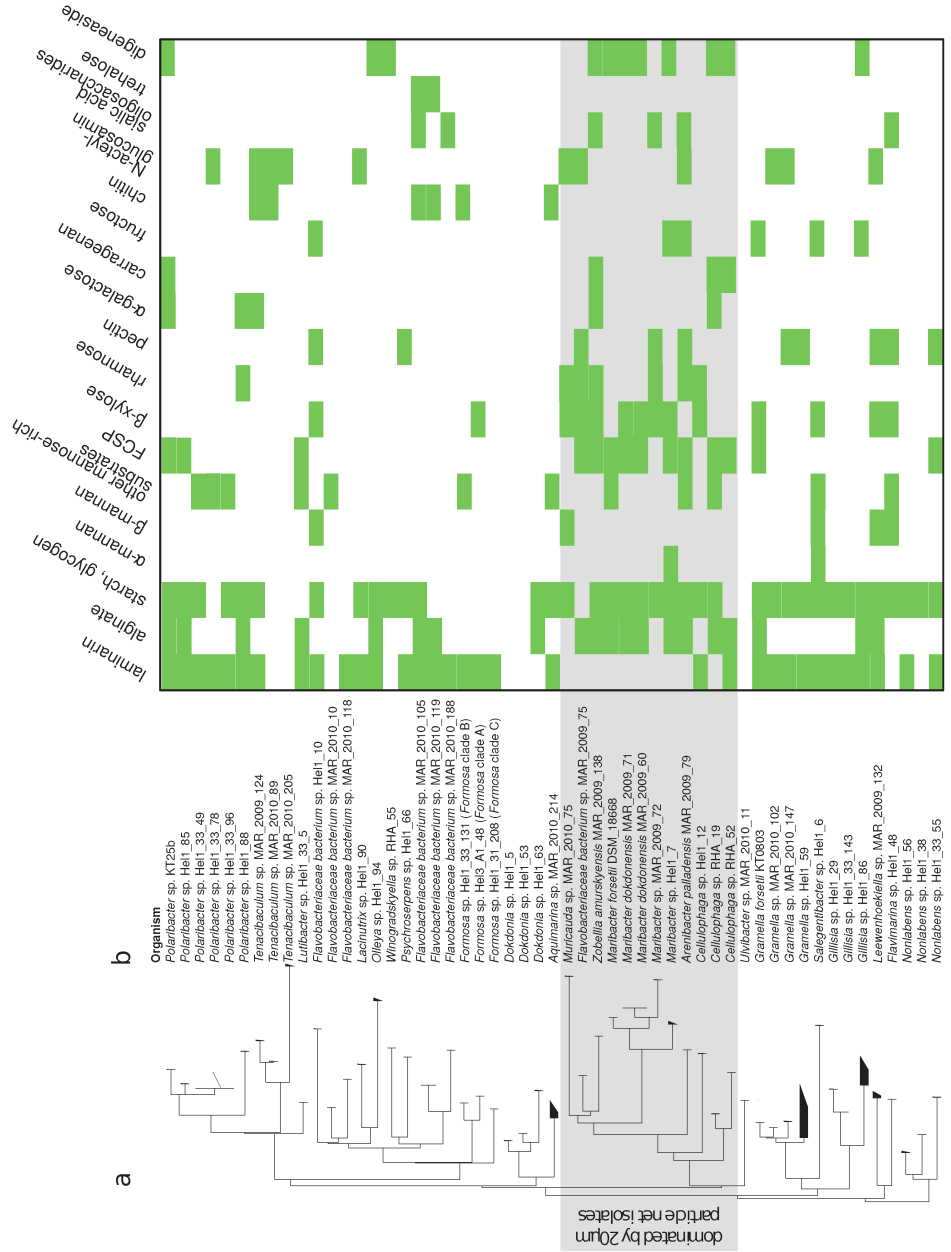


Figure 3.1: **a** Maximum-likelihood tree of 53 North Sea *Flavobacteria* isolates based on full-length 16S rRNA gene sequences. Scale bar: 10 nucleotide substitutions per 100 nucleotides. **b** Predicted degradation capacities of polysaccharide classes based on PUL-associated CAZyme annotations.

3.4.2 Putative substrate specificities

The 53 genomes revealed a wide range of as yet undescribed PULs. In total, 400 PULs were annotated, 259 of which could be linked to either dedicated polysaccharides or polysaccharide classes by in-depth annotations (Supplementary Table S3 - online).

Known and putative new laminarin PULs

Laminarins are β-1,3-linked glucans that are abundant as they act as storage compounds in brown algae and diatoms (both Stramenopiles). Forty-six PULs ($\sim$ 12%) were predicted to be involved in laminarin degradation, featuring four variants (Fig. 3.2): Variant A is a highly conserved, short PUL containing one predicted GH3 β-1,3-glucosidase framed by two predicted GH16 β-1,3(4)-glucanases (Fig. 3.2a). This arrangement was first described in *Gramella forsetii* KT0803 and shown to be upregulated by laminarin (Kabisch et al., 2014). Two distinct GH16 laminarinases have been studied in *Z. galactanivorans* DsiJ^{T}, which are endo-active (Labourel et al., 2014, 2015). For this species, it has been speculated that a cell surface-associated GH16 glucanase cleaves branched laminarin polysaccharides into oligosaccharides (Labourel et al., 2015), which can be transported through the SusC-like TBDT into the periplasm. Here, the GH3 β-1,3-glucosidase may further cleave off glucose units (Tamura et al., 2017), which are imported into the cytoplasm.

Variant B is a larger, more variable PUL (Fig. 3.2b). It shares homology with a PUL in *Polaribacter* sp. Hel1_33_49 that can be induced by laminarin (Xing et al., 2015). This PUL additionally features a predicted GH30 exo-β-1,6-glucanase and at least two GH17 β-1,3-glucan hydrolases with predicted endo- and exo-activities, respectively. The GH30 exo-β-1,6-glucanase removes β-1,6-glucose side chains from laminarin (Becker et al., 2017). Although GH16 enzymes can hydrolyze both β-1,3- and β-1,4-linked glucans, GH17 glucan hydrolases are highly specific to undecorated β-1,3 glucans and can have endo- (Reese & Mandels, 1959) and exo-activity (Barras & Stone, 1969). The β-1,3-glucan endohydrolase thus likely cleaves laminarin into oligosaccharides, which may be further degraded into glucose by the β-1,3-glucan exohydrolase.

Variants C and D PULs are likewise predicted to be capable of laminarin degradation based on gene content but have not been described before (Fig. 3.2c and d). They feature an additional putative GH5 glucan hydrolase with a carbohydrate-binding domain that binds β-1,3- and β-1,4-glucans (CBM6c, Michel et al. (2009)). They furthermore contain GH16 and GH30 family enzymes as described in variant B, but no GH17 enzymes.

In total, 62% (33/53) of all isolates and 78% (25/32) of surface water isolates contained at least one laminarin PUL. Variant A occurred 21 times, B 17 times, C five times, and D two times (Supplementary Table S3 - online). Eight isolates possessed two laminarin PULs (*Flavobacteriaceae bacterium* spp. MAR_2010_105 and MAR_2010_119, *Gramella* sp. MAR_2010_102, *Polaribacter* spp. Hel1_33_49, 78, 96 and Hel1_88 and *Psychroserpens* sp. Hel1_66) and three isolates contained three (*Formosa* spp. Hel3_A1_48 and Hel1_33_131, *Flavobacteriaceae bacterium* sp. Hel1_10). In contrast, laminarin PULs were far less prevalent in isolates obtained from the > 20 µm retentate (2/12). Laminarins are composed of a β-1,3-glucan backbone ramified by β-1,6 and, less frequently, β-1,2-linked glucose side chains (Gügi et al., 2015). The backbone length and ramification degree varies in different species. Laminarin of brown algae is capped at the reducing end by a 1-linked d-mannitol (Read et al., 1996). Only three isolates with laminarin PULs, namely the *Polaribacter* spp. Hel1_85 and KT25b and *Gramella* sp. MAR_2010_102, also possessed an annotated mannitol-2-dehydrogenase. It is possible that this enables utilization of brown algal laminarin. However, free mannitol is a more likely substrate. Growth on free mannitol has for example been demonstrated in the marine flavobacterium *Z. galactanivorans* (Groisillier et al., 2015). Studies on *Ectocarpus siliculosus* have shown that brown algae can store substantial amounts of free mannitol as compatible osmolyte (Gravot et al., 2010). Furthermore it has recently been shown that free mannitol is likewise frequently found in various planktonic microalgae (Tonon et al., 2017). Interestingly, diatoms seem to have lost their ability to synthesize mannitol, although exceptions exist (Tonon et al., 2017). The fact that phytoplankton blooms in the southern North Sea are usually diatom-dominated would hence explain, why mannitol-2-dehydrogenase genes were rarely found in our isolates. Consequently, the majority of isolates with laminarin PULs seem to only target diatom-type non-mannitol-capped chrysolaminarins, indicating that these are the major available laminarins in the southern North Sea.

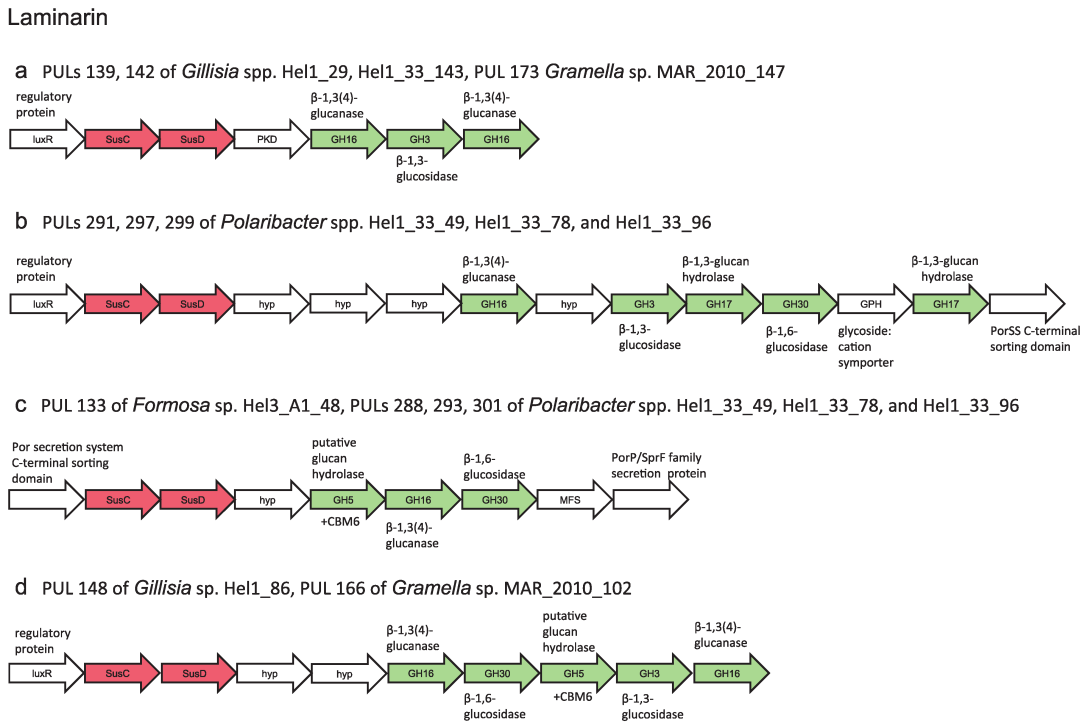


Figure 3.2: Conserved PULs known (**a**, **b**) and predicted (**c**, **d**) to target laminarin.

α-1,4-glucan (starch, glycogen)

PULs predicted to target α-1,4-glucans, such as starch, glycogen, and amylose were also highly abundant (43/400 PULs, 37/53 isolates; Supplementary Table S3 - online). Respective PULs often featured a *susCDE*-like gene triplet, at least one predicted GH13 α-glycosidase and frequently GH65 phosphorylases as well as GH31 hydrolases acting on α-glucosidic linkages (Fig. 3.3a). In some cases, these PULs also encoded a GH97 family glycoside hydrolase, known to hydrolyze diverse α-1,2-, α-1,3-, α-1,4-, and α-1,6-linked glycosidic bonds (Kitamura et al., 2008, Smith & Salyers, 1991). A similar PUL was first described for *Gramella forsetii* KT0803 and found to be upregulated in response to glucose-polymer substrates (Kabisch et al., 2014). The PUL depicted in Fig. 3.3a likely facilitates utilization of α-1,4-glucans featuring α-1,6-branches, such as the starch molecule amylopectin or potentially bacterial glycogen. Contrastingly, some isolates featured reduced versions of this PUL without any annotated *susE*- or *susF*-like gene and only one GH13 and GH65 gene, respectively (e.g., all *Maribacter* isolates). These isolates may only target simple non-branched α-1,4-glucans such as maltodextrin or amylose. Recent investigations have shown that *B. thetaiotaomicron* SusE is an immobile outer membrane protein that can modify the preferred sizes of maltooligosaccharides for uptake (Foley et al., 2018, Tuson et al., 2018). Hence SusE

homologs, whereas not essential, might be generally involved in fine-tuning the size selection of glycan uptake.

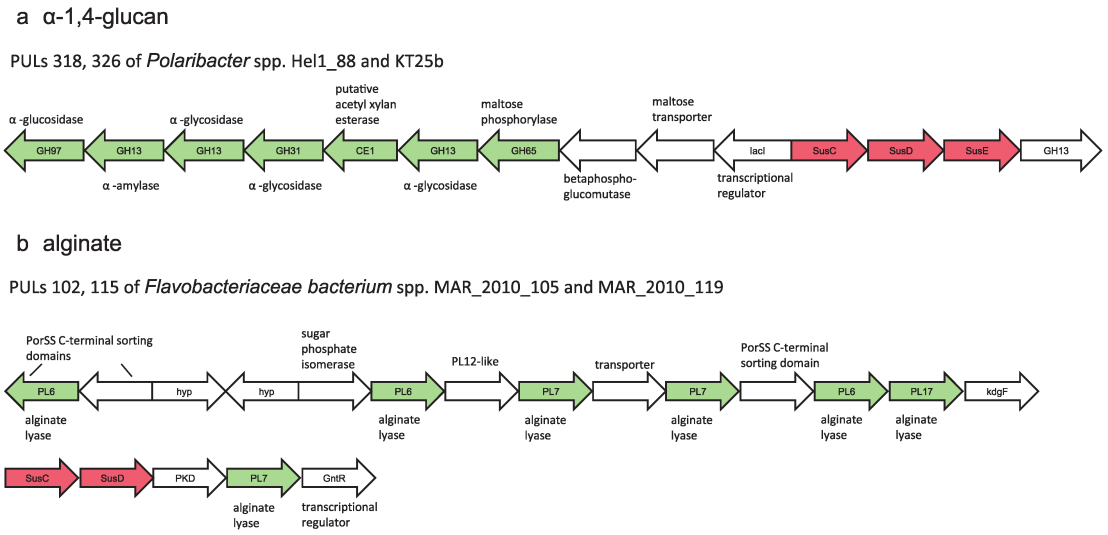


Figure 3.3: PULs predicted to target **a** α-1,4-glucans and **b** alginate.

Alginate

Twenty isolates featured in total 27 alginate-specific PULs (Fig. 3.1). Alginate consists of blocks of β-D-mannuronic acid and α-l-guluronic acid, forming a linear β-1,4-linked chain (Fischer & Dörfel, 1955). Corresponding PULs encode family PL6, 7, and 17 alginate lyases (Fig. 3.3b). Six of the alginate PULs also contained genes with sequence similarities to the sparsely investigated PL12 family. Known PL12 enzymes cleave heparin – a polymer of β-1,4-linked uronic acids and glucosamine that is often highly sulfated (Ulaganathan et al., 2017). Heparin and alginate hence are both linear, β-1,4-linked C5-uronans. However, sulfated alginates analogous to heparin, whereas being artificially synthesized for biotechnological uses (Arlov et al., 2014, Arlov & Skjåk-Bræk, 2017, Mhanna et al., 2014), have not been reported in nature. The latter is in line with a lack of sulfatases in the alginate PULs. Therefore, the PL12 family enzymes encoded in the alginate PULs likely represent novel types of alginate lyases. No PL15 and only one potential PL14 family alginate lyase (*Lutibacter* sp. Hel1_33_5, not PUL-associated) were annotated. The putative microalgae-associated cluster had a higher prevalence (8/12) of alginate PULs as compared with the other isolates (12/41) (Fig. 3.1).

Mannose-rich substrates

Seventeen isolates harbor PULs rich in mannose-targeting CAZyme genes, e.g., from families GH26, 38, 76, 92, 125, and 130 (Fig. 3.1, Supplementary Table S5 - online).

These PULs share little conservation in terms of gene arrangements, and only few contain a GH76 endo-α-1,6-mannanase (2/17) or a GH26 family endo-β-1,4-mannanase (5/17), indicative for either a linear α-1,6- or β-1,4-mannan backbone. Instead, GH92 and GH130 family genes are particularly prevalent. The GH92 family comprises solely exo-acting α-mannase genes. PULs rich in GH92 genes thus might target α-mannose-rich N-glycosylated glycoproteins that occur widespread in eukaryotes, including algae. GH130 enzymes comprise phosphorylases and glycoside hydrolases that act on β-mannosides and are known to partake in the degradation of β-mannans (Senoura et al., 2011).

The alpha-mannosidase-encoding PULs can be divided into two subtypes: (A) PULs containing multiple GH92 (α-1,2/3/4/6) genes that are often also rich in sulfatase genes (Fig. 3.4a; e.g., Supplementary Table S3 - online: PULs 289, 296, 300 of *Polaribacter* spp. Hel1_33_49/78/96; PUL126 of *Formosa* sp. Hel1_33_131), and (B) PULs with α-mannan-targeting CAZymes of diverse additional families, such as GH76 endo-α-1,6-mannanases, GH125 exo-α-1,6-mannosidases or GH38 α-mannosidases (α-1,2/3/6). These type (B) PULs are notably devoid of sulfatase-coding genes, indicating a non-sulfated substrate (Fig. 3.4b). A PUL with a similar CAZyme repertoire in *B. thetaiotaomicron* facilitates utilization of yeast cell wall α-mannan (Cuskin et al., 2015). Type (A) sulfatase- and GH92-rich PULs have been observed previously in *Polaribacter*-affiliated North Atlantic fosmids (Gómez-Pereira et al., 2012) and *Polaribacter* sp. Hel1_33_49 (Xing et al., 2015) and therefore seem to be widespread. In our case the PUL contains additional GH2, 3 and 88 family enzymes (Fig. 3.4a). Whereas GH families 2 and 3 are functionally diverse, GH88 enzymes are unsaturated β-glucuronyl hydrolases. This functional combination of CAZymes suggests degradation of α-glucomannans such as glucuronomannan, a polysaccharide that has been reported for diatoms (Ford & Percival, 1965, Gügi et al., 2015, Le Costaouëc et al., 2017) and brown algae (Wu et al., 2015) and thus should be abundant in the southern North Sea. Finally, co-located peptidases and a gene distantly related to GH99, a family that is reported to contain glycoprotein endo-α-mannosidases, indicate that this hypothesized glucuronomannan substrate might be a glycoprotein. However, functional studies are required to support this hypothesis. Le Costaouëc and colleagues (2017) recently revealed the main cell wall polysaccharide of the diatom *Phaeodactylum tricornutum* and possibly many other diatoms (Chiovitti et al., 2005) to be a 'linear poly-α-1–3-mannan decorated with sulfate ester groups and β-d-glucuronic residues'.

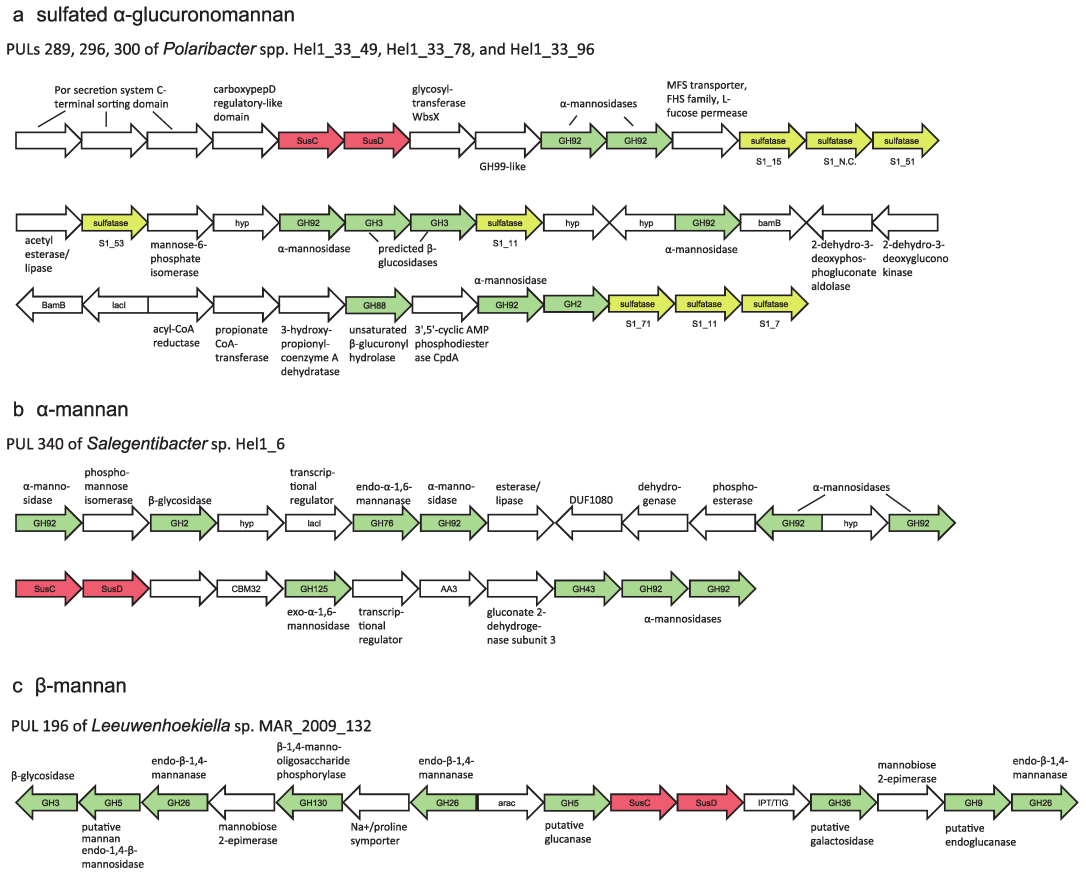


Figure 3.4: Selected PULs predicted to target mannose-rich substrates. Possible targets are **a** a sulfated α-mannan, **b** a non-sulfated α-mannan, and **c** a non-sulfated β-mannan. For sulfatases, families and sub-families are indicated below the genes.

The β-mannan PULs (Fig. 3.4c) all contained GH130 β-1,2(4)-mannooligosaccharide phosphorylases and GH26 CAZymes which are primarily composed of predicted endo-β-1,4-mannanases (Supplementary Table S3 - online: PUL 6 of *Flavobacteriaceae bacterium* sp. Hel1_10; PUL 196 of *Leeuwenhoekiella* sp. MAR_2009_132; PUL 211 of *Flavimarina* sp. Hel1_48; PUL266 of *Muricauda* sp. MAR_2010_75; PUL 342 of *Salegentibacter* sp. Hel1_6). Beta-mannans have been reported in the red macroalga *Porphyra umbilicalis* and in various species of the green macroalga *Codium*. Moreira & Filho (2008) proposed that 'in some algae species, linear (beta-) mannan seems to replace cellulose as the main cell wall glycan'.

Fucose-containing sulfated polysaccharides (FCSP)

Twenty PULs in 14 isolates suggest that FCSPs are also common substrates to marine *Flavobacteriia*. A prominent substrate of this group is fucoidan, a highly diverse polysaccharide prominent in brown algae. It contains l-fucose and sulfate ester groups owing

to its backbone of α-1,3 or alternating α-1,3/1,4-linked l-fucopyranosyl residues (Ale et al., 2011). This backbone has side chains containing diverse other monosaccharides, uronic acids, acetyl groups, and proteins (Li et al., 2008). In accordance with the structural complexity of FCSPs, PULs display equally complex gene repertoires, averaging 38 genes per PUL. A relatively short PUL is exemplarily shown in Fig. 3.5a. Characteristic CAZymes of predicted FCSP PULs in the isolated *Flavobacteriia* were GH29 and GH95 family α-l-fucosidases or potentially α-l-galactosidases. Other regularly co-occurring CAZymes included GH117 family enzymes (Supplementary Table S5 - online), β-xylosidases mostly of the family GH43, but also GH30, 39, and 120, and diverse α- and β-glucosidases of the families GH2, 3, 31, and 97. Sulfated FCSPs such as xylofucoglucans or -glucuronans have been reported for brown algal hemicelluloses (Kloareg & Quatrano, 1988, Popper et al., 2011) and might also occur in diatoms (Gügi et al., 2015).

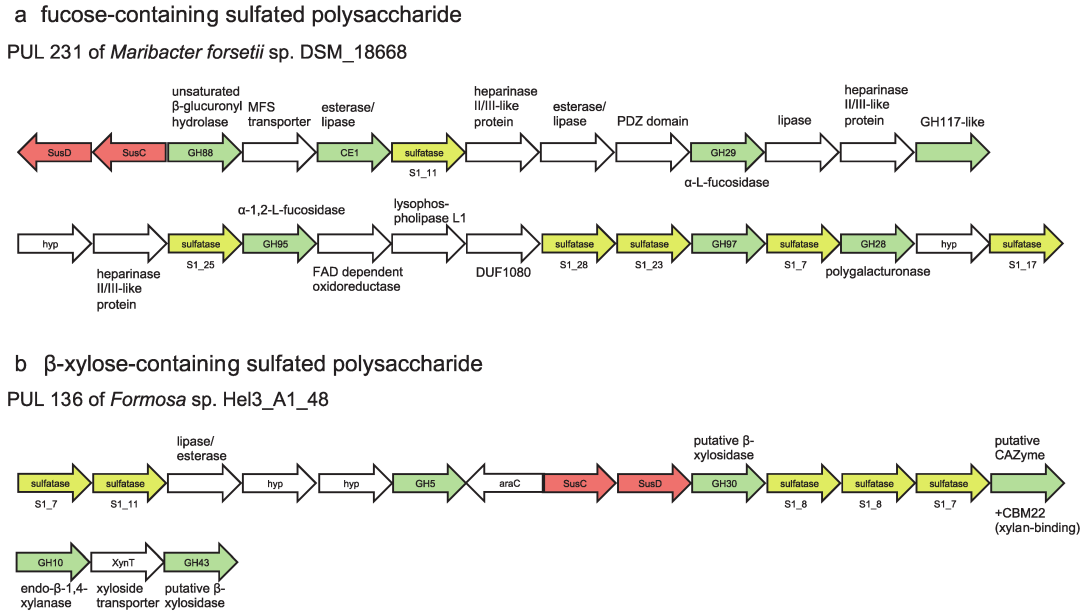


Figure 3.5: PULs predicted to target sulfated substrates rich in **a** fucose (FCSP) and **b** xylose. For sulfatases, families, and sub-families are indicated below the genes.

β-xylose-containing substrates

Twenty-two PULs predicted to target β-xylose-rich substrates were found in 14 isolates (Fig. 3.1, Supplementary Table S3 - online). Likely substrates are heterogeneous β-xylans such as arabinoxylans, glucuronoxylans, and sulfated xyloglucans. These PULs encode GH10 and GH43 enzymes targeting xylans, as well as GH30, 115, and 67 enzymes. GH10 enzymes are endo-β-1,4-xylanases capable of cleaving large β-1,4-xylan backbones into oligosaccharides. GH30 and GH43 enzymes have broader degradation capacities,

and while both families contain β-xylosidases, they were also reported to target mixed xylose-containing substrates such as arabinoxylan and α-l-arabinofuranosides (GH43) or even completely different substrates such as β-glucosylceramidase or β-1,6-glucanase (GH30). GH67 and 115 can cleave glucuronic acid side chains from native xylans and are present in four PULs that might target glucuronoxylans (PULs 243, 256 of *Maribacter* spp. Hel1_7, MAR_2009_72; PUL 269 of *Muricauda* sp. MAR_2010_75; PUL 338 of *Salegentibacter* sp. Hel1_6; Supplementary Table S3 - online). Two PULs were predicted to target arabinoxylans through a GH51 α-l-arabinofuranosidase (PUL 155 of *Gramella forsetii* KT0803; PUL 199 of *Flavimarina* sp. Hel1_48; Supplementary Table S3 - online). Four PULs predicted to target β-xylose-rich substrates encode sulfatases (PUL136 of *Formosa* sp. Hel3_A1_48, Fig. 3.5b; PULs 363, 364, 366 of *Zobellia amurskyensis* MAR_2009_138; Supplementary Table S3 - online). Marine xylans have been reported as hemicellulose components in green (Sørensen et al., 2011), red and brown macroalgae (Kloareg & Quatrano, 1988, Painter, 1983), and as cell wall components in some diatoms (Murray et al., 2007, Wustman et al., 1998).

Further substrates

Further possible substrates comprised sulfated α-rhamnose- and α-galactose-containing substrates, pectin, arabinan, trehalose-like α-1,1-glucans, N-acetylglucosamine and its polymer chitin, digeneaside, fructose, and sialic acid-containing polysaccharides. These compounds are discussed in the supplementary text.

3.4.3 Trees of SusC- and SusD-like proteins reveal substrate-specific clusters

We computed trees for all SusC- and SusD-like protein sequences of the 400 isolate PULs and obtained pronounced clusters for many of the predicted polysaccharide substrates (Fig. 3.6). For clarity, functionally heterogeneous or undefined clusters are depicted as gray triangles (complete trees: Supplementary Figures S1A, S1B - online). Well-defined clusters in both trees included the structurally simple polysaccharides laminarin, α-1,4-glucans and alginate. For example, SusD-like proteins of laminarin-targeting PULs of *Cellulophaga* sp. RHA_52, *Flavobacteriaceae bacterium* sp. Hel1_10, *Formosa* sp. Hel1_33_131 and *Psychroserpens* sp. Hel1_66 (PULs 58, 71, 128, 331, Supplementary Table S3 - online), exhibited between 64 and 78% identity (Supplementary Table S4 - online), whereas identity to SusD-like sequences from other PULs within the same respective genome was only 10–25% (data not shown).

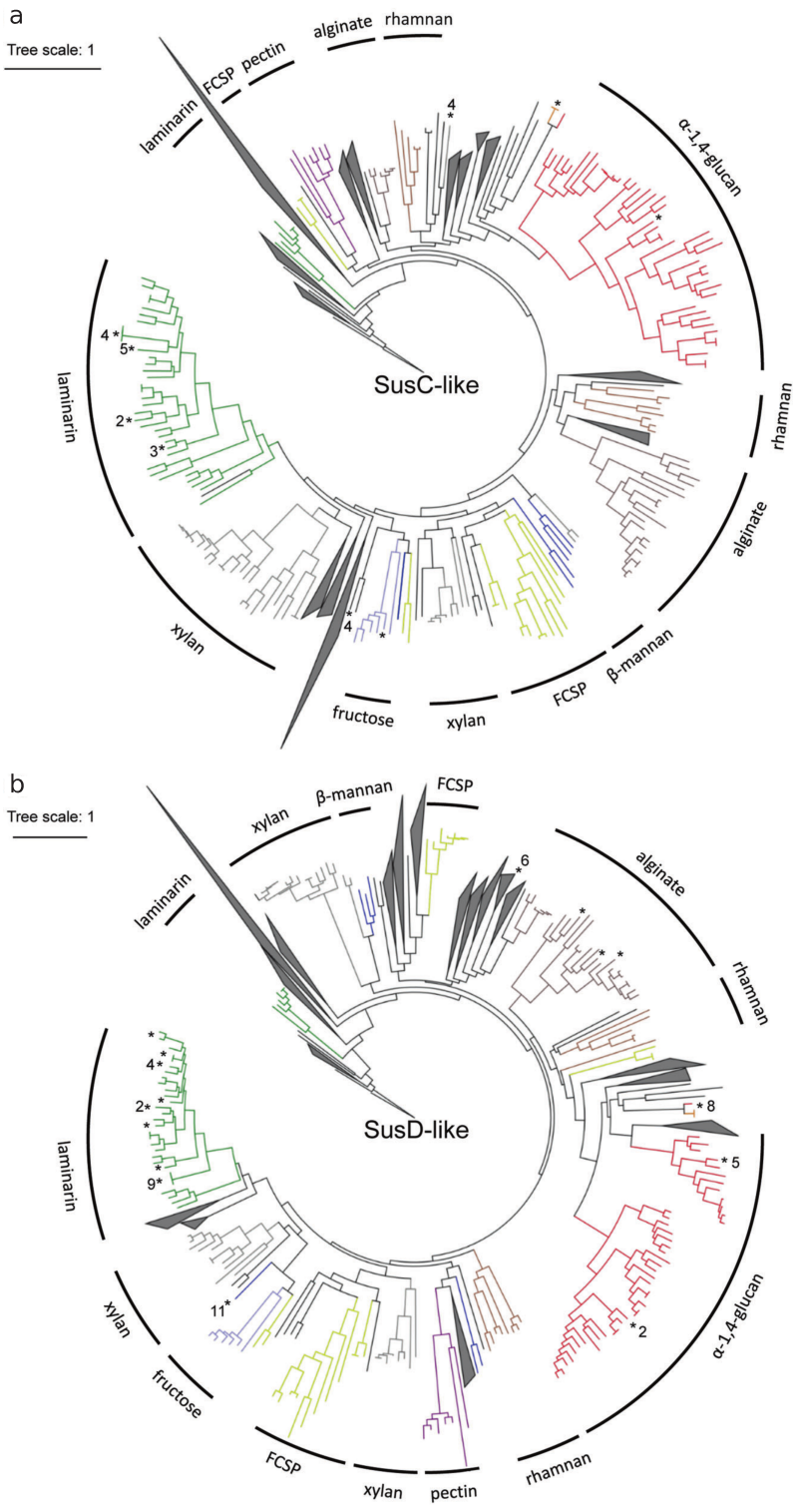


Figure 3.6: Trees of all PUL-associated SusC- (**a**) and SusD-like (**b**) proteins of the *Flavobacteriia* isolates showing functional, substrate-specific clustering. Protein sequences were aligned using the MAFFT G-INS-i algorithm and trees were calculated using FastTree 2.1.5 approximate-maximum likelihood (SusC-like: 370; SusD-like: 362). Substrate predictions are depicted in colors. Proteins with expressed homologs in North Sea bacterioplankton blooms of more than 40% sequence identity are marked with asterisks (and number of homologs if x > 1). Corresponding figures labeled with protein sequence identifiers, originating species and PUL-associated CAZymes are provided as supplementary material.

SusC/D-like proteins from conserved PULs for these structurally simple substrates were more closely related than those from more variable PULs targeting structurally more diverse substrate classes such as FCSPs or xylose-rich substrates (Supplementary Table S4 - online). This is visible in the trees by shorter and longer respective branch lengths (Fig. 3.6). Some substrates formed multiple clusters, for example xylose-rich substrates. This might indicate either rather different xylose-containing substrates or multiple ways of attack and uptake for a given class of xylose-containing substrate.

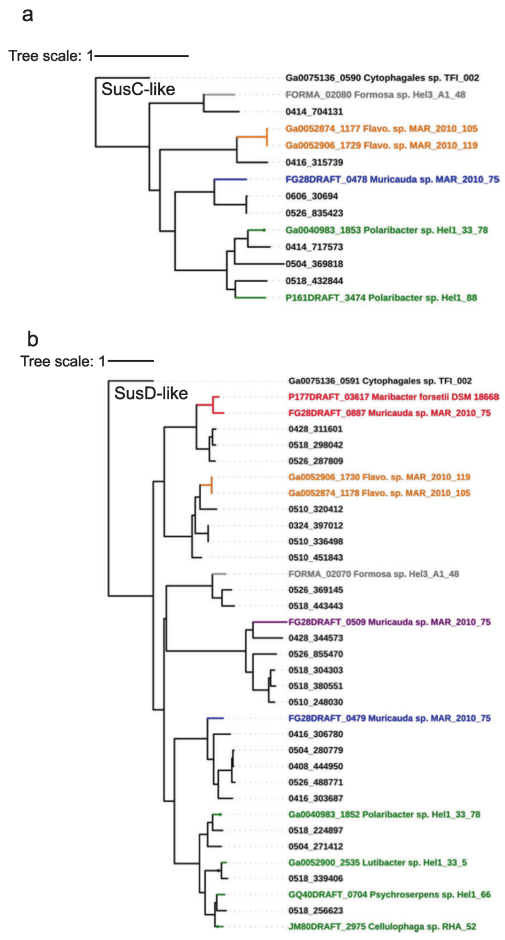
The topologies of the SusC- and SusD-like protein trees were notably congruent regarding branching patterns of the identified substrate-specific clusters. Only the pectin cluster was located at a distinctly different position. SusC- and SusD-like proteins from the same PULs exhibited a strong tendency to occur in corresponding substrate-specific clusters in both trees. This applied to $> 70\%$ of the SusC and SusD sequences within identified substrate-specific clusters (Supplementary Figures S1A and S1B - online).

3.4.4 SusC/D-like protein expression of bacterioplankton during phytoplankton blooms supports temporal variations of polysaccharide abundances in situ

SusC/D-like proteins range among the highest expressed proteins in bacterioplankton metaproteomes from productive oceans (Morris et al., 2010, Teeling et al., 2012, Williams et al., 2013). Likewise, studies on flavobacterial isolates have identified SusC/D-like proteins as the highest expressed proteins within PULs that are furthermore co-regulated with other PUL-encoded proteins including CAZymes (Kabisch et al., 2014, Xing et al., 2015). SusC/D expression thus represents a suitable proxy for overall PUL expression

We monitored bacterioplankton spring phytoplankton blooms in the southern North Sea during 2009 with weekly, and in 2010 to 2012 with about monthly sampling (Teeling et al., 2012, 2016). At 14 selected time points we analyzed the free-living 0.2–3 µm bacterioplankton using shotgun metaproteomics (total: 23,917 identified proteins), and detected high numbers of expressed SusC/D-like proteins in metaproteomes across all sampled years (Supplementary Table S2 - online).

To identify potential substrates, we aligned all expressed SusC/D-like sequences (SusC: 390; SusD: 118) to the SusC/D-tree constructed from isolate PULs. Isolate sequences with highest similarities ($\geq 40\%$) to expressed sequences are indicated in Fig. 3.6. Further semi-quantitative analyses were confined to SusC/D-like proteins where at least one related homolog reached expression levels of ≥ 0.05 %NSAF, i.e. 0.05% of all mass-adjusted spectral counts (see Materials and methods; Fig. 3.7).



c

%NSAF (=SAF/sum all SAFs*100)

	2009				2010				2011			2012		
PP bloom state	MID	MID	MID	LATE	PRE	EARLY MID	LATE		PRE	MID	LATE	PRE	EARLY	LATE
date / substrate	03.31.	04.07.	04.14.	04.21.	03.03.	04.08.	05.04.	05.18.	03.24.	04.28.	05.26.	03.08.	04.16.	05.10.
sulfated xylan	0.00	0.00	0.08	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
α-1,1-glucan	0.05	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.19	0.01	0.00	0.00	0.02
β-mannan	0.04	0.05	0.00	0.02	0.00	0.00	0.09	0.05	0.00	0.22	0.00	0.00	0.00	0.00
	0.03	0.04	0.02	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00
laminarin	0.00	0.04	0.07	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.01
	0.00	0.00	0.00	0.00	0.00	0.00	0.13	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	0.08	0.05	0.01	0.01	0.00	0.03	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00

d

α-1,4-glucan	0.01	0.04	0.01	0.02	0.00	0.11	0.06	0.02	0.00	0.01	0.01	0.00	0.00	0.01
	0.02	0.01	0.00	0.01	0.00	0.05	0.01	0.01	0.00	0.07	0.00	0.00	0.00	0.01
	0.01	0.01	0.00	0.00	0.00	0.00	0.06	0.06	0.00	0.00	0.03	0.00	0.00	0.00
α-1,1-glucan	0.00	0.00	0.01	0.03	0.00	0.01	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.04	0.07	0.08	0.03	0.04	0.01
	0.03	0.08	0.06	0.12	0.00	0.03	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.02	0.01	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.08	0.00	0.00	0.00	0.03
sulfated xylan	0.00	0.00	0.00	0.07	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.03	0.02	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
arabinan	0.01	0.02	0.02	0.06	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.09	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.03	0.00	0.00	0.05	0.00	0.00	0.01
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.07
β-mannan	0.04	0.02	0.05	0.09	0.00	0.02	0.01	0.02	0.00	0.03	0.01	0.00	0.01	0.02
	0.11	0.09	0.03	0.04	0.00	0.07	0.11	0.02	0.00	0.25	0.01	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.15	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.01	0.01	0.00	0.00	0.00	0.00	0.00	0.01	0.00	0.00	0.02	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.06	0.01	0.00
laminarin	0.00	0.01	0.00	0.00	0.00	0.00	0.00	0.04	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.01	0.01	0.00	0.00	0.00	0.07	0.02	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.00	0.00	0.00	0.02	0.02	0.00	0.00	0.00	0.00	0.00	0.00	0.00
	0.00	0.00	0.01	0.03	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00	0.00

Figure 3.7: **a**, **b** Trees of expressed SusC and SusD-like proteins identified in 3–0.2 µm bacterioplankton during North Sea spring phytoplankton blooms in 2009–2012 using proteomics. The most closely related SusC/D-like sequences from North Sea *Flavobacteriia* isolates in this study were integrated in the tree. Protein names correspond to sequence identifier and isolate name. Sequences were aligned using the MAFFT G-INS-i algorithm. The tree was calculated using FastTree 2.1.5 approximate-maximum likelihood. **c**, **d** Corresponding expression levels as Normalized Spectral Abundance Factors (%NSAF) for the four consecutive blooms. Metaproteomic samples were classified as pre-, early-, mid-, and late-bloom based on chlorophyll a concentrations during the spring phytoplankton blooms. Expression levels are highlighted by green color.

Laminarin

Homologs to laminarin-binding SusC-like proteins were detected amidst the 2009 and 2010 phytoplankton blooms, with one homolog reaching a notable maximum of 0.13 %NSAF on May 4th, 2010 (Fig. 3.7b). Respective SusD-like homologs were detected in the same years and highest expression was observed at the same date in 2010 (0.07 %NSAF, Fig. 3.7d). Amino-acid identities of expressed SusC homologs and laminarin PUL SusC-like proteins from isolates ranged from 48–68%, and for SusD homologs from 40–78% (Supplementary Table S4 - online). Our data suggest that laminarin occurred at the bloom peaks in 2009 and 2010 and directly thereafter. This is supported by detection

of expressed GH3 β-glucosidases and GH16 β-glucanases in 2009 (Teeling et al., 2012) and, to a lesser degree, in 2010 (Supplementary Table S2 - online). Chrysolaminarin is produced by microalgae such as *Thalassiosira nordenskioeldii* diatoms or representatives of *Phaecystis haptophytes* (Alderkamp et al., 2007a, Myklestad, 1989), which both were among the dominating microalgae in 2009 and 2010 (Teeling et al., 2016).

Alpha-1,4-glucan

Respective SusD-like proteins were most abundantly detected in 2010, peaking on April 8th (0.11 %NSAF, Fig. 3.7d), but also in 2009 and 2011. Sequence identities to isolate α-1,4-glucan PUL SusD-like proteins ranged from 43 to 46% (Supplementary Table S4 - online). These data indicate that α-1,4-glucans, potentially starch or glycogen, represented a recurring substrate from 2009 to 2011 during early to late phytoplankton bloom stages.

Alpha-1,1-glucan

An α-1,1-glucan-binding SusC-like protein potentially targeting a trehalose-like substrate was strongly expressed on April 28th 2011 (0.19 %NSAF; Fig. 3.7b), but also in all other years except 2010. Its sequence identity to the SusC-like proteins of the trehalose PULs of *Flavobacteriaceae bacterium* spp. MAR_2010_105 and MAR_2010_119 was 43% (Supplementary Figure S2J - online, PUL103 and PUL116, Supplementary Tables S3 and S4 - online). Corresponding SusD-like proteins were detected in all years, but most strongly throughout the blooms of 2009 and 2011. Their protein identities to the isolate PULs ranged from 58–59% (Supplementary Table S4 - online).

Sulfated β-xylan

One SusC and two SusD-like proteins likely targeting a sulfated β-xylan were expressed in the mid and late stages of the phytoplankton bloom of 2009, peaking at 0.08 %NSAF for SusC-like proteins and 0.07 %NSAF for SusD-like proteins. Their identities to homologs of the sulfated β-xylan PUL of *Formosa* sp. Hel3_A1_48 (Fig. 3.5b, PUL136, Supplementary Table S3 - online) was 53% and 49–53%, respectively (Supplementary Table S4 - online).

Beta-mannan

Homologs with high identities to SusC/D-like proteins occurring in a predicted β-mannan PUL from *Muricauda* sp. MAR_2010_75 were strongly expressed during blooms from 2009 to 2011, peaking on April 28th 2011 for SusC (0.22 %NSAF) and SusD (0.25 %NSAF) homologs. No SusC-like proteins of putative β-mannan PULs were detected in 2012 and SusD-like expression was likewise much weaker. The expressed SusC-like proteins were 52–60% identical to the ones from the β-mannan PUL of *Muricauda* sp. MAR_2010_75 (PUL266, Supplementary Table S3 - online) and the SusD-like proteins showed 45–51% identity (Supplementary Table S4 - online). The predicted β-mannan PUL of *Muricauda* sp. MAR_2010_75 harbors two pairs of SusC/D-like proteins. The one with expressed in situ homologs did not cluster with those from other beta-mannan PULs in our SusC/D trees. Hence the two SusC/D-like pairs might target different oligosaccharides. As some PULs can be induced by substrates other than those that they degrade, it is possible that the substrate that led to the upregulation of the in situ homologs was not a beta-mannan. Proteomic studies of this PUL in *Muricauda* sp. MAR_2010_75 are required to clarify regulation of this PUL and to interpret the in situ data.

Arabinan

SusD-like proteins potentially targeting an arabinan were expressed at late phytoplankton blooms stages during all four years with at least 0.05 %NSAF. However, their identities to the SusD-like protein of a predicted arabinan PUL from *Muricauda* sp. MAR_2010_75 were only 26–30% (Supplementary Figure S2I - online, PUL267, Supplementary Tables S3 and S4 - online).

In summary, comparative analyses of SusC/D homolog expression are indicative of a successive utilization of different polysaccharides over the course of phytoplankton blooms. This agrees with successive changes in the microbial community composition during bloom events that we reported earlier on (Teeling et al., 2012, 2016).

3.5 Discussion

PUL function predictions in this study are based on sequence similarity analyses and thus cannot rival time-consuming laboratory-based functional studies in terms of accuracy. Knowledge on polysaccharides from marine algae, in particular from microalgae, is still

sparse and thus false predictions are possible. Still, the holistic approach to analyze the PUL spectrum of a large number of isolates from a single habitat allows identification of recurrent and thus important PULs as targets for future functional studies and to build testable hypotheses on possible substrates.

We observed diverse polysaccharide degradation capacities among North Sea *Flavobacteriia* with no distinct correlation to taxonomy. Even isolates from identical genera often featured notably diverging PUL repertoires and genome sizes (e.g., *Polaribacter*, *Maribacter*, and *Cellulophaga*), substantiating earlier data (Xing et al., 2015). Our findings suggest that a species' PUL repertoire is more dependent on its distinct ecological niche, whereas its phylogeny is of secondary importance. This corroborates the hypothesis that PULs are exchanged between *Flavobacteriia* through horizontal gene transfer (Hehemann et al., 2012b).

The isolates' PUL repertoires showcase that abundant, structurally simple substrates such as laminarin, α-1,4-glucans, and alginate are targeted by likewise conserved and frequent PULs. These substrates are likely so common that preserving the respective catabolic machinery is favorable for many marine *Flavobacteriia*. Diatom-derived chrysolaminarin has been estimated to amount to 5–15 petagrams of organic carbon annually (Alderkamp et al., 2007b) and accordingly laminarin-specific PULs were frequent in our surface water isolates. The four predicted laminarin PUL variants we identified might indicate that different laminarin types (Gügi et al., 2015) are targeted by different PULs or that some of these PULs act as helper modules in laminarin degradation, as many species feature more than one laminarin PUL type (e.g., *Formosa* spp. Hel1_33_131 and Hel3_A1_48, *Gramella* sp. MAR_2010_102). Variant B contains predicted endo- and exo-acting β-1,3-glucan hydrolases (GH17) highly specific to laminarin degradation (Becker et al., 2017). Variants A, C, and D only contain GH16 endo-1,3(4)-β-glucanases and may not be restricted to laminarin, but are potentially capable of degrading further mixed-linkage β-1,3/1,4-glucans, as recently shown for a similar conserved PUL in human gut *Bacteroidetes* (Tamura et al., 2017). Clustering of the SusC/D sequences of variants A, B and D in the SusC/D trees support that they bind the same substrate (Supplementary Figure S1A, S1B - online). Those of variant C, however, are located elsewhere, indicating that this PUL might indeed have an alternate function. Functional studies on model strains containing variant C (e.g., *Formosa* sp. Hel3_A1_48) and D (e.g., *Gramella* sp. MAR_2010_102) will be necessary to ultimately elucidate the functions of these PULs.

Alginate and α-1,4-glucan degradation capacities were prevalent in the isolates obtained from the > 20 µm retentate, which might be microalgae-associated, but were also common in many seawater isolates. Overall, laminarin, α-1,4-glucan, and alginate PULs are

fairly conserved and make up over a quarter of all PULs in the isolates (115/400), suggesting that these are abundant polysaccharide substrates in North Sea coastal habitats that many microbes can consume and likely compete for.

Other substrates are utilized by fewer isolates, which implies that algal polysaccharide degradation is usually carried out by multiple resource-partitioning bacterioplankton species. In putative microalgae-associated isolates, these substrates include new FCSP variants, and xylose- and rhamnose-rich polysaccharides. Among surface water isolates, these substrates are sulfated α-mannans (likely an α-glucuronomannan glycoprotein), β-mannans, sulfated α-galactans and β-xylans, chitin, and (trehalose-like) α-1,1-glucans.

A major result of this study is the substrate-specific clustering of both SusC- and SusD-like proteins. The strong tendencies of SusC and SusD homologs to occur in corresponding substrate-specific clusters in both trees, resulting in similar tree topologies, suggest coevolution of these two proteins. This hypothesis is corroborated by recent X-ray crystallography findings showing complex formation of two SusC- and SusD-like proteins of *B. thetaiotaomicron* (Glenwright et al., 2017). Clustering was more pronounced for structurally conserved, simple polysaccharides than for the heterogeneous and partially new substrates described in this study. This is expected, as heterogeneous substrates are attacked at multiple points resulting in a variety of structurally different oligosaccharides for uptake. Furthermore, broad substrate classes that currently can only be defined as, e.g., FCSPs or xylose-containing substrates might actually represent multiple chemically rather different substrates. Hence, improvement of functional clustering is to be expected once more detailed knowledge on algal polysaccharides structures is available.

We here provide first metaproteomic data indicating that high-resolution expression analysis of SusC/D homologs may be used for monitoring changes in microbial polysaccharide degradation activity. This provides a proxy on which polysaccharides are important at a given time and space in marine carbon cycling. Considering our still incomplete knowledge, only expressed SusC/D homologs exhibiting a high level of sequence identity to functionally annotated or characterized SusC/D sequences should be considered. Absence of such expressed homologs, however, does not preclude that a respective substrate may be targeted by an as yet unknown SusC/D system. This current limitation notwithstanding, our approach provides a new method to identify environmentally relevant polysaccharide substrates that due to their structural complexity are still difficult to identify by direct chemical analysis.

3.6 Acknowledgements

We thank Sabine Kühn and Ingrid Kunze for cultivation and DNA extraction and Bernhard Fuchs for critical reading. Genome sequencing and assembly was conducted in the framework of the COGITO project by the US Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, and is supported by the Office of Science of the US Department of Energy under Contract No. DE-AC02-05CH11231. This study was funded by the Max Planck Society and supported by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the research unit FOR2406 'Proteogenomics of Marine Polysaccharide Utilization (POMPU)' by grants of Rudolf Amann (AM 73/9-1), Hanno Teeling (TE 813/2-1), and Thomas Schweder (SCHW 595/10-1) and the DFG Emmy Noether program (Jan-Hendrik Hehemann grant HE 7217/1-1). Lennart Kappelmann and Karen Krüger are members of the International Max Planck Research School of Marine Microbiology (MarMic).

Chapter 4

During marine microalgae blooms few *Bacteroidetes* clades mediate the bulk of bacteroidetal remineralization of algal glycans using a restricted set of genes

Karen Krüger[1], Meghan Chafee[1], T. Ben Francis[1], Tijana Glavina del Rio[2], Dörte Becher[3], Thomas Schweder[4,5], Rudolf I. Amann[1], Hanno Teeling[1]

[1]Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany
[2]DOE Joint Genome Institute, 2800 Mitchell Drive, Walnut Creek, CA 94598, USA
[3]Institute for Microbiology, University Greifswald, Felix-Hausdorff-Straße 8, 17489 Greifswald, Germany
[4]Pharmaceutical Biotechnology, Institute of Pharmacy, University Greifswald, Felix-Hausdorff-Straße 3, 17487 Greifswald, Germany
[5]Institute of Marine Biotechnology, Walther-Rathenau-Straße 49a, 17489 Greifswald, Germany

In preparation for *The ISME Journal*.

Contributions to the manuscript:
Experimental concept and design: 65%
Acquisition of experimental data: 20%
Data analysis and interpretation: 90%
Preparation of figures and tables: 100%
Drafting of the manuscript: 95%

Electronic figure versions and supplementary tables are available at https://owncloud.mpi-bremen.de/index.php/s/vYhAlReQrNHh1Xp.

4.1 Abstract

Bacteroidetes recurrently thrive during phytoplankton blooms as they are adapted for the degradation of phytoplankton-derived organic matter such as polysaccharides. We investigated bacterioplankton during spring algae blooms in the southern North Sea in 2010 to 2012 using a time-series of 38 deeply sequenced metagenomes. Automatic binning yielded 6 455 metagenome-assembled genomes (MAGs), from which we extracted 1 286 refined *Bacteroidetes* MAGs covering ∼120 mostly uncultivated species. We identified thirteen dominant, recurrent *Bacteroidetes* clades carrying a restricted set of conserved polysaccharide utilization loci that likely mediate the bulk of bacteroidetal algal polysaccharide degradation. The majority of loci were predicted to target the diatom storage polysaccharide laminarin, alpha-glucans, alpha-mannan-rich substrates and sulfated xylans. Analyses of key-players' abundances and PUL repertoires suggested that fewer and simpler polysaccharides dominated early bloom stages, and that more complex polysaccharides became available as blooms progressed.

4.2 Introduction

Spring and summer blooms of planktonic microalgae (phytoplankton) are annually recurring phenomena in the world's oceans temperate coastal regions. Such blooms can vary in overall phytoplankton composition, and likewise the composition within an individual bloom can undergo multiple successive composition changes over its course. These dynamics notwithstanding, 16S ribosomal RNA gene-based studies have revealed recurring diversity patterns within the community of free-living bacteria (bacterioplankton) that respond to phytoplankton blooms (Andersson et al., 2010, Chafee et al., 2017, Fuhrman et al., 2015, Gilbert et al., 2012, Needham et al., 2018, Teeling et al., 2016). Heterotrophic members of the *Bacteroidetes*, *Gammaproteobacteria* and the alphaproteobacterial *Roseobacter* clade are often among the dominant and recurring responders that partake in the degradation of marine phytoplankton-derived organic matter (Buchan et al., 2014).

A substantial portion of phytoplankton organic matter consists of polysaccharides that act as storage compounds, cell wall components and exudates. The proportion of these polysaccharides varies depending on algae species and growth stage, and ranges from 13% to about 90% of algal dry-mass (Myklestad, 1974). Monosaccharide compositions are often dominated by glucose, mannose, fucose, arabinose, xylose, rhamnose and galactose (reviewed in Gügi et al. (2015) and Mühlenbruch et al. (2018)). However, linkage types and the overall glycan structures remain mostly elusive as so far only few structures

have been resolved (de Jesus Raposo et al., 2014, Gügi et al., 2015, Le Costaouëc et al., 2017). One of the better studied glycans is laminarin that acts as storage compound in brown algae and diatoms (both stramenopiles). Annual production has been estimated to amount to 5 to 15 Pg (1 Pg x a^{-1} = 10^{15} g x a^{-1}; Alderkamp et al. (2007b)), which is equivalent to a third of the world's oceans total annual primary production of 45 to 50 Pg (Field et al., 1998).

Enzymes for the degradation and modification of polysaccharides are widespread among bacterial and archaeal phyla. *Bacteroidetes* evolved a unique degradation machinery that is usually encoded in so called polysaccharide utilization loci (PULs; Bjursell et al. (2006)). The characteristic feature of a polysaccharide utilization locus (PUL) is a gene tandem coding for a SusC-like TonB-dependent transporter and a SusD-like protein. Co-located are genes that encode degradative Carbohydrate-Active enZymes (CAZymes), such as glycoside hydrolases (GHs), polysaccharide lyases (PLs) or carbohydrate esterases (CEs), and accessory proteins, often including proteases and sulfatases. In a concerted effort, extracellular CAZymes degrade polysaccharides into size-ranges that can pass via the TonB-dependent transporter into the periplasm, where they are protected from competing bacteria and can be further degraded to monosaccharides. These monosaccharides are subsequently taken up across the cytoplasmic membrane by means of dedicated transporters. TonB-dependent uptake is either selfish by taking up all initial cleavage products from the cell surrounding and thereby avoiding cross feeding, as described for α-mannan degradation by a human gut *Bacteroidetes* (Cuskin et al., 2015), or semi-selfish with some cleavage products available to other bacteria (Rakoff-Nahoum et al., 2016). Uptake of polysaccharides into the periplasmic space was recently demonstrated for marine *Bacteroidetes* using fluorescent labeling (Reintjes et al., 2017).

The first PUL was identified in the human gut bacterium *Bacteroides thetaiotaomicron* (Reeves et al., 1997) and acts on starch. Ever since PULs with diverse specificities have been described in bacteria from various habitats (for review see Grondin et al. (2017)), including metagenomic studies of the fecal microbiome of the North American beaver (Armstrong et al., 2018) or the moose rumen (Svartström et al., 2017). In strains of marine *Bacteroidetes* PULs have been experimentally verified that target agar/porphyran (Hehemann et al., 2010), alginate (Thomas et al., 2012), laminarin and alginate (Kabisch et al., 2014), laminarin (Xing et al., 2015), carrageenan (Ficko-Blean et al., 2017) and ulvan (Reisky et al., 2018, Salinas & French, 2017). These studies focused on selected model bacteria, whereas ecological studies that provide a more holistic perspective in terms of environmental relevance are just emerging (e.g. Bennke et al. (2013), Gómez-Pereira et al. (2012), Kappelmann et al. (2018)).

Using a series of ten metagenomes from spring phytoplankton blooms in the southern North Sea during the years 2010 to 2012 we could previously identify fingerprint-like patterns in CAZyme, transporter and sulfatase gene contents within distinct clades of free-living *Bacteroidetes* (Teeling et al., 2016). We also investigated PUL diversity by sequencing the genomes of 53 isolated strains of North Sea *Flavobacteriia* (Kappelmann et al., 2018). However, spacing of the metagenome time series was too wide to capture the rapid changes that were occurring within the bacterioplankton communities (Teeling et al., 2016). Isolated *Flavobacteriia* strains on the other hand rarely matched those that were abundant during the investigated algae spring blooms (e.g. Unfried et al. (2018)).

Recent advances in binning of metagenomes into metagenome-assembled genomes (MAGs; e.g. Hugerth et al. (2015), Parks et al. (2017)) bear the potential to substantially enhance our understanding of the PUL contents of as yet uncultured bacterioplankton species and thereby their contributions to the remineralization of phytoplankton-derived biomass (Delmont et al., 2015, Francis et al., 2018, Hugerth et al., 2015). This approach has so far not been applied to marine bacterial communities. We therefore added 28 metagenomes to our initial dataset, resulting in a dense series of 38 points in time for the consecutive spring blooms of 2010 to 2012. Automatic binning yielded 6 455 MAGs, from which we obtained 1 286 refined *Bacteroidetes* MAGs representing approximately 120 distinct species. This dataset provides genome data of so far unmatched temporal and taxonomic resolution of as yet uncultured algae bloom-associated marine bacterioplankton species.

4.3 Material and methods

4.3.1 Sampling

Surface seawater from the long-term ecological research site Kabeltonne at Helgoland island in the southern North Sea (54° 11.03′ N, 7° 54.0′ E) was collected during spring phytoplankton blooms from 2010 to 2012 as described elsewhere (Teeling et al., 2016). Free-living bacteria were separated from particle-attached bacteria by pre-filtration through 10 and 3 µm pore-size filters and collected on 0.2 µm pore-size filters. DNA was subsequently extracted from retained biomass. Corresponding metaproteomes were obtained during spring phytoplankton blooms from 2010 to 2012 as described in (Teeling et al., 2012) and (Kappelmann et al., 2018).

4.3.2 Metagenome sequencing, assembly and automated binning

38 surface seawater metagenome samples were sequenced at the Department of Energy Joint Genome Institute (DOE-JGI, Walnut Creek, CA, USA) as previously described in (Teeling et al., 2016), and quality filtering and trimming of raw reads, metagenome assembly and binning to metagenome-assembled genomes (MAGs) using CONCOCT (Alneberg et al., 2014) was performed as described previously (Francis et al., 2018). Details are provided in the supplementary material. Metagenome accession numbers are provided in Table S1 - online and MAGs have been submitted to ENA under accession PRJEB28156.

4.3.3 Phylogenomic analysis, bin refinement and reduction of redundancy

CheckM v1.0.7 (Parks et al., 2015) was used to assess MAG quality and to place MAGs in a reference genome tree using the lineage_wf workflow. The tree_qa command was subsequently used to extract taxonomic classification of MAGs. MAGs were selected for refinement, if (i) CheckM placed them within the *Bacteroidetes* and their completeness was >40%, or (ii) if at least one *susD*-like gene was annotated, irrelevant of taxonomic placement and completeness values. These criteria were met by 1 185 out of 6 455 automatically binned CONCOCT MAGs. The former were refined using the anvi'o (Eren et al., 2015) anvi-refine command through visually inspecting the coverage and GC profiles of contigs in each MAG and their positioning after hierarchical clustering. During this process contigs were removed if considered contamination. MAGs were split into several refined MAGs when profiles suggested so. This resulted in a set of 1 456 manually refined MAGs, 1 286 of which were classified as *Bacteroidetes* (verified by a second analysis with CheckM).

Since all 38 metagenomes were assembled and binned seperately, redundant MAGs were obtained from different sampling dates. To reduce this redundancy we used Mash v1.1.1 (Ondov et al., 2016) with default sketch size of 1 000 to cluster MAGs into 110 approximate species clusters (henceforth referred to as Mash-clusters). Mash estimates distances between genomes and calculated Mash distances have been shown to correlate well with average nucleotide identities (ANI), i.e. a Mash distance of $\leq$0.05 correlates with $\geq$95% ANI, a common average nucleotide identity threshold for genomes to be considered belonging to the same species (Goris et al., 2007). Mash-clusters were visualized using Cytoscape v3.5.0 (Shannon et al., 2003).

Mash-clusters with less than two MAGs $\geq$70% completeness (19 of 110) and singleton MAGs <70% completeness and contamination >10% (90 of 119), as evaluated by

CheckM using the *Bacteroidetes*-specific marker set, were excluded from further analysis (Table S2 - online). Two representative MAGs of each of the remaining 91 Mash-clusters, and the 29 remaining singleton MAGs (∼120 species) were analyzed using GTDB$_{tk}$ v0.0.8 with GTDB version 83 (Parks et al., 2018). GTDB$_{tk}$ uses a genome-based, manually curated taxonomy in which taxonomic ranks are normalized and polyphyletic groups are removed. Only closely related sequences to our MAGs were extracted from the reference alignment and retreed using RAxML v8.2.10 (Stamatakis, 2014) with automatic selection of substitution model and rapid-bootstrapping with 1 000 resamples (-m PROTGAMMAAUTO -p 12345 -x 12345 -# 1000) and subsequently visualized using iTOL (Letunic & Bork, 2016).

4.3.4 PUL prediction and SusC/D protein trees

Genes were predicted using Prodigal (Hyatt et al., 2010) from within anvi'o. PUL genes were predicted using a combination of hidden Markov models and BLAST (Altschul et al., 1990) searches against the CAZy database (Lombard et al., 2014). For HMMer a combination of hidden Markov models was used including the dbCAN (Yin et al., 2012) models for all CAZy families, the Pfam profile for sulfatases (PF00884), the SusD Pfam models (PF07980, PF12741, PF14322, PF12771) as also used by PUL-DB (Terrapon et al., 2018, 2015) and the TIGRFAM (Selengut et al., 2007) profile for SusC (TIGR04056; OMP_RagA_SusC). HMMer results were filtered with the db-CAN hmmer-scan-parser script, filtering for multiple annotations, e-values and a 30% domain coverage minimum. In a second step the HMMer results of all CAZy families were filtered based on whether these proteins had a DIAMOND BLAST hit (Buchfink et al., 2014) with ≥30% identity, at least 40% query coverage and an e-value of ≤E-20 against proteins from the same CAZy family in the CAZy database as of 2017/07/20 (downloaded from the dbCAN webpage). Only these annotations were considered for subsequent PUL prediction.

Potential PULs were extracted by finding all loci, where at least three predicted PUL genes (sulfatases, CAZymes, SusC and SusD proteins) were within close proximity (less than ten genes in between), unless it exclusively contained glycosyltransferases. Further processing required a PUL to have at least one *susC* or *susD* gene and at least two degradative CAZymes from the GH or PL families.

SusC and SusD protein sequences from all predicted PULs (SusC: 1 195; SusD: 1 311) were used for tree calculation. Included were PUL SusC and SusD proteins from isolate genomes (Kappelmann et al., 2018) and metaproteome SusC/Ds that were extracted from the metaproteome data (see details provided in supplementary material). The tree

was calculated using MAFFT v7.313 (Katoh & Standley, 2013) with L-INS-I for protein alignment and FastTree v2.1.10 (Price et al., 2010) for tree calculation.

Representative SusC/D proteins were selected from the tree, if the identical SusC or SusD (at least ≥95% nucleotide identity to closely clustering proteins) from a SusC/D pair was assembled and detected in a PUL from at least four metagenome time-points. In case of a close metaproteome SusC/D sequence (≥90% amino acid identity), three identical SusC/D proteins were considered sufficient. The selected SusC and SusD representatives (SusC: 131; SusD: 130), representative for in total 910 and 987 SusC and SusD proteins, were used for calculation of a reduced tree using the same SusC/D alignment, but RAxML v8.2.10 with automatic substitution model selection and rapid-bootstrapping with 1 000 resamples (-m PROTGAMMAAUTO -p 12345 -x 12345 -# 1000). Trees were visualized using iTOL.

Finally all representative SusC/Ds and respective PULs were linked to individual Mash-clusters and thus taxonomy. Species-level taxonomic affiliation of SusC and SusD proteins were considered as consistent, if half or more of the harboring contigs were binned into MAGs from the same Mash-cluster. In cases where less than half, but the majority of contigs were associated with the same Mash-cluster, PULs were considered as putative PULs for a Mash-cluster. Clade-level taxonomic affiliations were used tor those PULs that were binned into Mash-clusters of the same clade.

4.3.5 MAG and PUL abundance estimates

SPAdes error-corrected reads from all sampling dates were mapped onto singleton MAGs and onto the two representative MAGs of each Mash-cluster. In the latter case, mean values of the representatives were used as approximation for the Mash-cluster's abundance. Read mapping and post-filtering of SAM files was performed using the parameters described for metagenome sequencing, assembly and automated binning (supplementary material). Final abundance values were calculated as reads per kilobase million [RPKM = (number_of_mapped_reads_on_MAG * 1 000 000) / (length_of_MAG_in_kbp * total_number_of_reads)].

4.3.6 Metaproteome sequencing and availability

Metaproteome samples were obtained and processed as described previously (Teeling et al., 2012) with modifications described in (Kappelmann et al., 2018). Proteome mass spectrometry data have been deposited at the PRIDE database (PXD008238, 10.6019/PXD008238).

4.4 Results

4.4.1 Phylogeny of *Bacteroidetes* MAGs

Manual refinement of the initial 6 455 automatically generated MAGs yielded 1 286 MAGs affiliating with the *Bacteroidetes* phylum. GTDB$_{tk}$ analysis (Parks et al., 2018) revealed that about 75% of these belonged to clades that we previously identified as key players during North Sea spring blooms of 2010 to 2012 (Chafee et al., 2017, Teeling et al., 2016), including the *Formosa* (GTDB$_{tk}$ UBA3537), *Polaribacter*, *Aurantivirga* (Aurantivirga SCGC-AA160-P02), *Cd.* Prosiliicoccus (GTDB$_{tk}$ HC6-5), the NS3a and NS5 marine groups (GTDB$_{tk}$ MAG-120531 and MS024-2A) and MAG-121220-bin8, which all belong to the *Flavobacteriaceae* family (Fig. 4.1A, Fig. 4.7). Other families represented were the *Cryomorphaceae* including the VIS6 clade (GTDB$_{tk}$ UBA10364), the families UA16, 1G12, and *Crocinitomicaceae* (previously part of the *Cryomorphaceae* family), and the *Cyclobacteriaceae* of the *Cytophagales* order (Fig. 4.1A, Fig. 4.7). In total 27 multi-MAG Mash-clusters from these 13 *Bacteroidetes* clades were identified as having abundances exceeding five RPKM at a single time-point or 38 RPKM at all time-points combined (2 RPKM $\approx$ 1% relative abundance detected by fluorescence *in situ* hybridization (Teeling et al., 2016)).

In previous studies we isolated strains of the abundant genera *Formosa* and *Polaribacter* from the southern North Sea (Hahnke et al., 2015, Hahnke & Harder, 2013). Members of these genera reached well above 20% relative abundance during the North Sea spring phytoplankton bloom in 2009 (Teeling et al., 2012) and recurred in 2010 to 2012 (Teeling et al., 2016). We analyzed two isolated strains from each genus in functional studies (Unfried et al., 2018, Xing et al., 2015). Of these, *Polaribacter* sp. Hel1_33_49 and *Formosa* sp. Hel1_33_131 belong to the same species as Mash-cluster 55 and 47, respectively. The latter, however, was not among the abundant Mash-clusters, only reaching a maximum of 3.5 RPKM. For all other abundant Mash-clusters representing approximate species there are as yet no isolated strains.

MAG sizes within the 13 most abundant *Bacteroidetes* clades, were mostly in the range of 1.5 to 3 Mbp, which conforms to the lower range of the *Bacteroidetes* genome size spectrum (e.g. Xing et al. (2015)). In particular MAGs from the NS5 marine group, the MAG-121220-bin8 genus, the *Cryomorphaceae* and two unclassified *Flavobacteriaceae* clades were consistently below 2 Mbp (Fig. 4.1B, Fig. 4.7). MAGs from the closely related genera *Aurantivirga* and *Polaribacter* exhibited the broadest size spectra, ranging from 1.7 to 3.5 Mbp and 2.1 to 4.3 Mbp, respectively (Fig. 4.1B, Fig. 4.7).

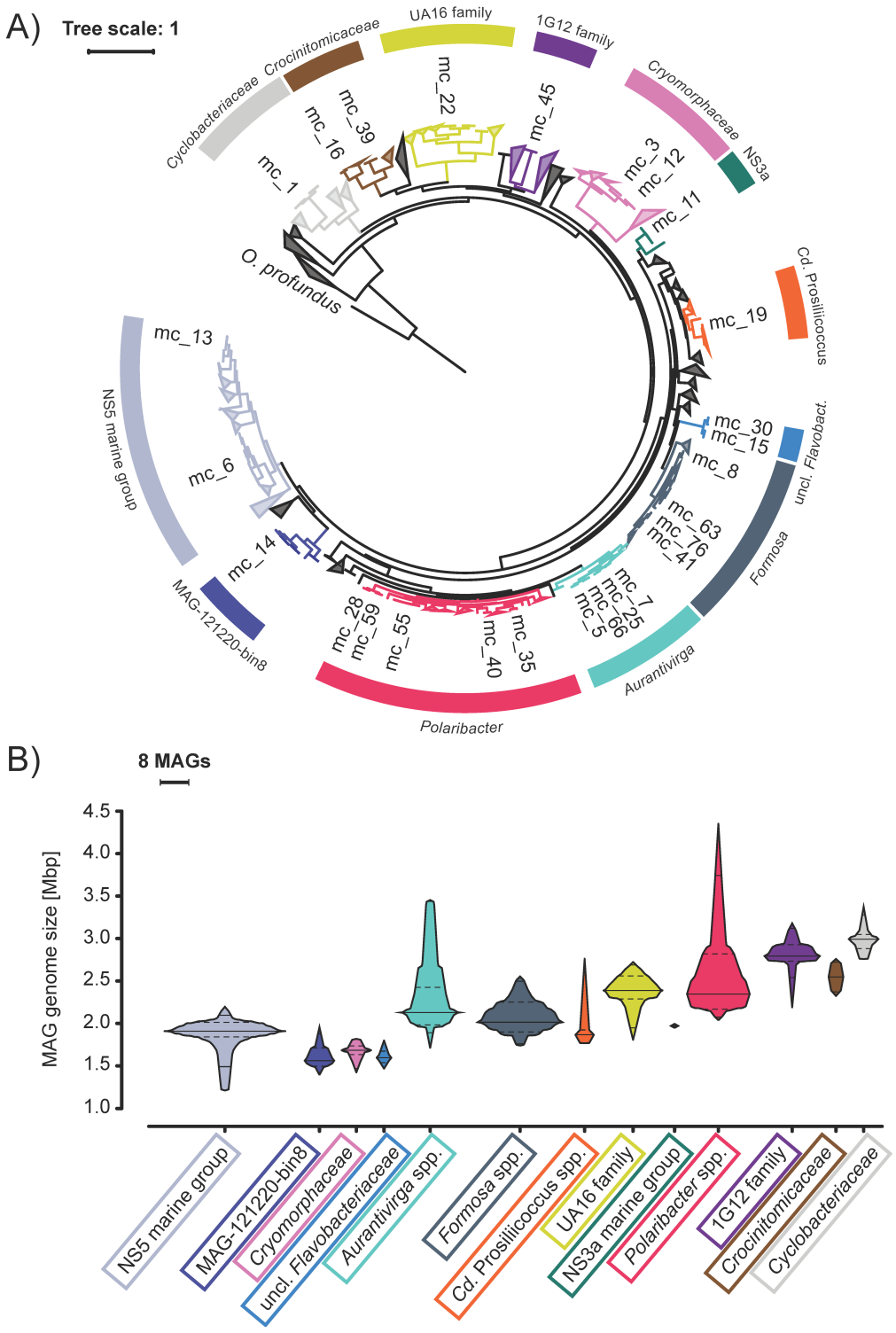


Figure 4.1: (A) Maximum-likelihood tree of Mash-clusters based on concatenated marker proteins according to the GTDB$_{tk}$ genome phylogeny (Parks et al., 2018). Only Mash-clusters (mc) were included that contained at least two MAGs ≥70% completeness, plus additional 29 singleton MAGs with ≥70% completeness and ≤10% contamination. Mash-clusters reaching high abundances (>5 RPKM at one time-point or 38 RPKM at all time-points combines) during the sampling period are highlighted together with their taxonomic affiliations. Scale bar: mean number of amino acid substitutions per site. Outgroup: *Oceanithermus profundus*. (B) Violin plots of MAG size distributions within clades of abundant Mash-clusters (colored areas). Area width corresponds to MAG numbers. Solid lines represent median and deciles, and dashed lines represent quartiles.

4.4.2 Seasonality of *Bacteroidetes* MAGs

Based on time-point and chlorophyll *a* concentration we categorized our 2010 to 2012 metagenomes into pre-bloom, (mid) bloom and post-bloom seasons (Fig. 4.8). The bloom periods of 2010 and 2011 metagenomes were further subdivided into primary and secondary blooms, since blooms in these years were bimodal with two extended chlorophyll *a* peaks. Phytoplankton composition differed between spring blooms, with 2010 and 2011 being dominated by different diatom species and in 2012 a less intense and shorter bloom dominated by *Chattonella* species (Teeling et al., 2016). Nonetheless MAGs of individual Mash-clusters often originated from all three or at least two years (Fig. 4.2A, 4.9). This corroborates that a larger number of abundant species recurred, as we suggested previously (Chafee et al., 2017, Teeling et al., 2016). A few Mash-clusters contained MAGs from almost all sampling time-points, indicating that the corresponding species are autochthonous rather than specialized on bloom situations. In contrast, Mash-clusters of *Formosa* spp., *Polaribacter* spp., *Cd.* Prosiliicoccus spp. as well as Mash-clusters 6, 25 and 30 displayed a specifically high prevalence during bloom and post-bloom stages (Fig. 4.2B).

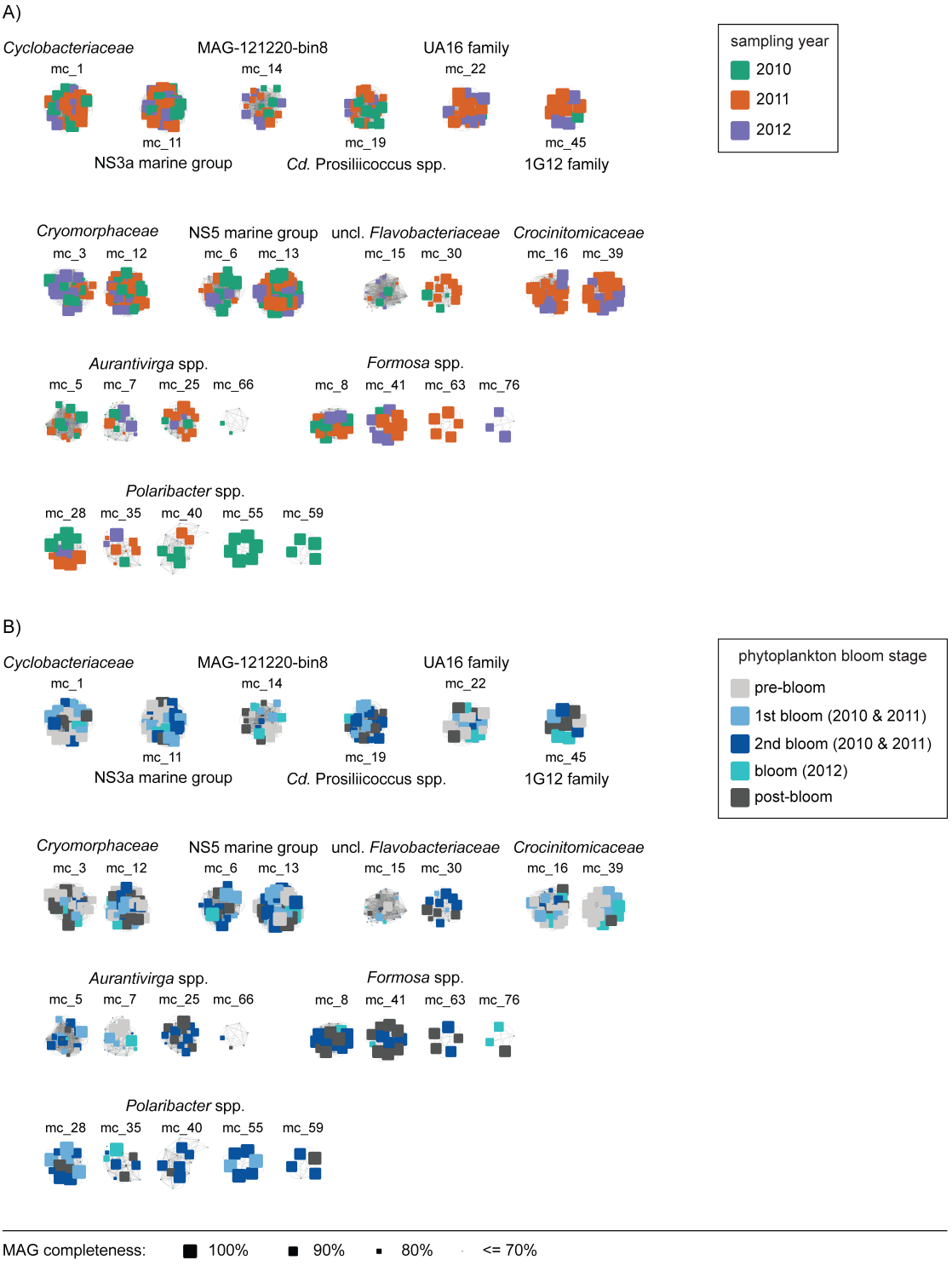


Figure 4.2: Composition of abundant *Bacteroidetes* Mash-clusters with respect to sampling year (A) and phytoplankton bloom stage (B). Squares represent individual MAGs with sizes corresponding to completeness, while gray lines indicate Mash distances $\leq$0.05, and thus the approximate species connections for MAGs in each Mash-cluster. (A) Color coding indicates the year in which individual MAGs were retrieved. The majority of Mash-clusters are multi-year clusters as they contain MAGs assembled in all three or at least two years. (B) Color coding represents phytoplankton bloom stages, revealing Mash-clusters predominantly retrieved from metagenomes obtained during phytoplankton blooms, such as e.g. *Formosa* spp., *Polaribacter* spp. and *Cd.* Prosiliicoccus spp. Mash-clusters.

4.4.3 PULs in *Bacteroidetes* Mash-clusters

We identified 117 representative PUL-containing contigs with a total of 131 *susC* and 130 *susD* genes. Some contigs carried two or three *susC/D* pairs or even two distinct PULs. 78% of these *susC/D* pairs could be assigned to specific Mash-clusters and 17% putatively to specific Mash-clusters or a specific genus. Genome trees of corresponding protein sequences showed clustering based on predicted PUL substrate specificity rather than a taxonomic signal, as we have previously described (Kappelmann et al., 2018). Substrate predictions suggested five major polysaccharide categories that free-living *Bacteroidetes* targeted during the investigated North Sea spring phytoplankton blooms (Fig. 4.3).

Beta-glucans/laminarin

Predicted β-glucan or laminarin PULs were found for many of the abundant *Bacteroidetes* genera. In the SusC/D trees most of the laminarin PULs clustered in a single clade with few exceptions clustering separately, e.g. the β-glucan PULs from the NS5 marine group (Fig. 4.3). The main cluster comprised 37 laminarin PULs that attributed to at least ten genera. These PULs could be divided into three variants, all of which carried a GH16 gene (Fig. 4.4A-C). The shortest variant showed a combination of GH3 and GH16 enzymes, sometimes accompanied by a GH2 or a GH30_1 (Fig. 4.4A; variant A in (Kappelmann et al., 2018)). This variant was present in several Mash-clusters attributed to the *Formosa*, *Cd.* Prosiliicoccus and *Polaribacter* genus, Mash-clusters 15 and 30 (uncl. *Flavobacteriaceae*), and the families UA16, 1G12 and *Cryomorphaceae*.

A second variant with the combination of GH16, GH30_1 and often two GH17 enzymes (Fig. 4.4B, variant B in (Kappelmann et al., 2018)) was mainly present in *Polaribacter* and *Aurantivirga*. Other less abundant genera carrying this variant were Mash-clusters 9 (*Algibacter*-related), 10 (genus UBA6710), 34 (genus UBA1994) and 70 (unclassified on genus level).

The third variant contained a combination of a CBM6-containing GH5_46 and a GH16 enzyme (variant C in (Kappelmann et al., 2018)). This PUL type has so far not been experimentally verified as targeting laminarin, but consistent clustering of the respective SusC/D sequences with verified laminarin PULs suggests laminarin or β-glucans as substrates. Automatically predicted PULs from the PULDB often show these enzymes in combination with either a GH30 or a GH3. This PUL type in the large laminarin cluster was restricted to the VIS6 *Cryomorphaceae* and the UA16 genus within the UA16 family.

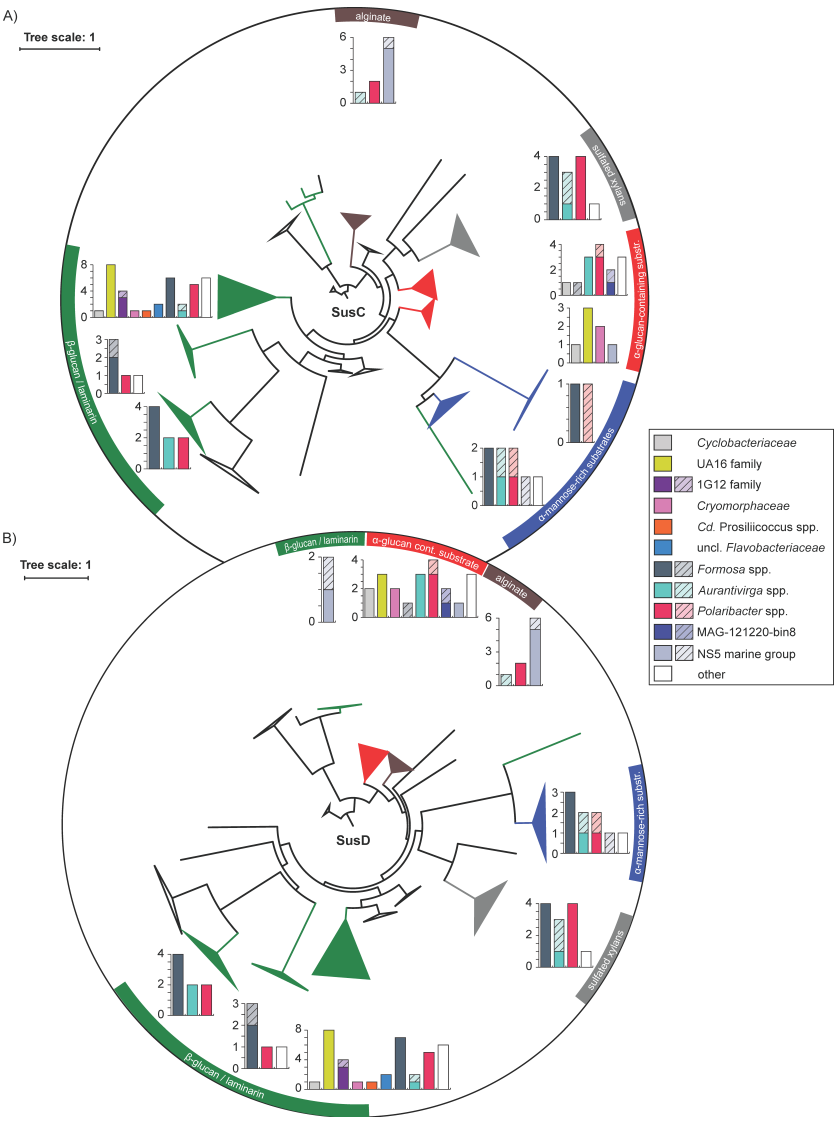


Figure 4.3: SusC (A) and SusD (B) tree of all 131 representative SusC-like and 130 SusD-like proteins located within PULs in the metagenome dataset. Protein sequences were aligned using MAFFT L-INS-I (Katoh & Standley, 2013) and trees calculated using RAxML (Stamatakis, 2014). Branches were collapsed and color coded based on substrate predictions derived from CAZyme analysis of the respective PULs (Kappelmann et al., 2018). Bar plots indicate taxonomic affiliation of respective PUL contigs for each collapsed branch. Block colors indicate SusC and SusD proteins whose corresponding contigs were consistently binned into a single Mash-cluster, and hatched areas indicate binning consistent on genus level.

PULs similar to this third variant were present in separate clusters in the SusC/D tree, and belonged mostly to *Formosa*, *Polaribacter* or *Aurantivirga* Mash-clusters. These PULs code for a combination of GH5_46, GH30_1 and GH16, though the GH16 was missing in some of the *Formosa* PULs. At least three such PULs were located on contigs with two *susC/D* pairs (Fig. 4.4C) - one pair of which clustered in the large laminarin cluster. To clarify whether these PULs are co-regulated, target different laminarin types or increase each others efficiency can only be tested experimentally with isolate strains carrying such a PUL combination.

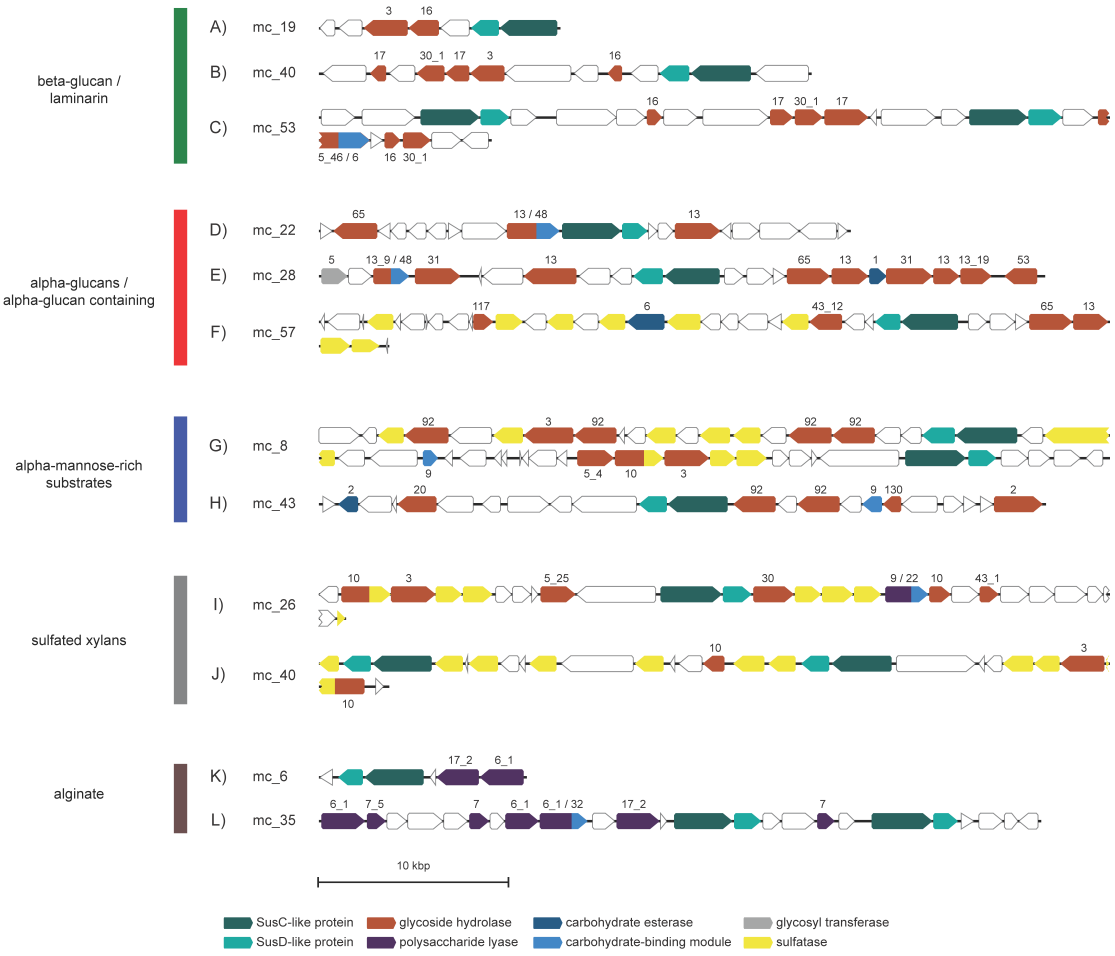


Figure 4.4: Polysaccharide utilization loci representing common detected PUL patterns. Mash-cluster affiliations for each PUL-contig are indicated. Color coding of genes indicate gene types and corresponding numbers indicate CAZyme family associations.

Alpha-glucan-containing substrates

Larger clusters were also observed for PULs related to the degradation of α-glucans. The SusD tree contained one and the SusC tree two such clusters. Except for two, the corresponding PULs were characterized by GH13 family genes, which are known to act on α-glucose. These PULs can be separated into rather simple variants containing maximal four to five CAZymes and more complex and variable PULs, some of which contain sulfatases. The simpler PULs were prevalent within the VIS6 *Cryomorphaceae* and the UA16 families and contained a GH65 and one to two GH13 genes (Fig. 4.4D). The latter sometimes had a CBM48, a carbohydrate-binding module known to bind α-glucans. The GH13 plus GH65 combination has already been described in a PUL from *Gramella forsetii* KT0803^{T} as targeting α-1,4-glucans (Kabisch et al., 2014). The GH13 family, among other functions, contains α-amylases that act on amylose of starch or glycogen. The GH65 family contains maltose phosphorylases that cleave maltose to glucose and glucose-1-phosphate.

The more complex PULs carried several GH13 and often a GH65 gene, additionally often a GH43_12 gene and sometimes sulfatases (Fig. 4.4E-F). The substrate of these PULs remains unclear, but the similarity in SusC and SusD protein sequences and the presence of GH13 enzymes suggest that the oligosaccharide taken up is similar in structure compared to those originating from the initial breakdown of simple α-glucans. In terms of taxonomy these more complex PULs could be detected in clades such as the *Cyclobacteriaceae*, the NS5 marine group, *Formosa* and the MAG-121220-bin8 genus.

Alpha-mannose-rich substrates

SusC and SusD sequences from GH92-rich PULs also formed coherent clusters. GH92 family proteins are exo-acting α-mannosidases that cleave different linkage types (e.g. Zhu et al. (2010)). Some but not all of these PULs also contained a larger number of sulfatase genes.

Sulfatase-containing GH92-rich PULs harbored three to four GH92 genes with up to six sulfatase genes and additional glycoside hydrolases of families GH2, GH3 and GH43_2 (Fig. 4.4G). These PULs affiliated with *Formosa*, *Polaribacter*, *Aurantivirga* and an *Algibacter*-related Mash-cluster (mc_9). Similar types have already been reported for *Polaribacter* and *Formosa* species (Gómez-Pereira et al., 2012, Kappelmann et al., 2018, Xing et al., 2015) and were hypothesized to target glucuronomannan substrates e.g. the cell wall polysaccharide of the diatom *Phaeodactylum tricornutum* (Kappelmann et al., 2018), which has been described as a linear poly-α-1,3-mannan decorated with sulfate ester and β-D-glucuronic residues (Le Costaouëc et al., 2017).

Sulfatase-free GH92-rich PULs contained a combination of GH92, GH130, GH20, GH18 genes and sometimes a CBM9-containing, GH3 or GH2 gene (Fig. 4.4H). Often a predicted CE2 was also present. Similar PULs have been predicted in *Flavobacterium* sp. SCGC AAA160-P02 (*Aurantivirga*) and various *Salegentibacter* species in the PUL-DB. It is noteworthy that similar PULs have been detected on *Flavobacteriia* fosmids from the North Atlantic, which have been hypothesized to target mixed-linkage glucans including enzymes involved in xylan metabolism (CE2 acetyl-xylan esterase and CBM9) and degradation of mannose-rich substrates (GH92 and GH130) (Bennke et al., 2016). The sulfatase-free GH92 PULs in our study were distributed among *Formosa*, *Polaribacter* and *Aurantivirga* genera and the NS5 marine group.

Sulfated xylans/xylose-rich substrates

PULs putatively targeting sulfated xylans or xylose-rich substrates showed two conserved patterns. The first included a sulfatase domain-containing GH10, a GH3 and two additional sulfatase genes (Fig. 4.4I-J). The second harbored a GH30, three sulfatases, a PL9 with a CBM22, a GH10 and a GH43_1 gene (Fig. 4.4I). These two patterns were sometimes combined in a single PUL. Known GH10 enzymes are mostly endo-β-xylanases. Families GH3, 30 and 43_1 have a broader substrate spectrum, but all include β-xylosidases. These two types affiliated with *Formosa*, *Aurantivirga* and *Polaribacter* MAGs and have already been described to be present in marine *Bacteroidetes* (Bennke et al., 2016, Kappelmann et al., 2018).

Alginate

The alginate PUL cluster was dominated by NS5 marine group Mash-clusters, though also two *Polaribacter* and one *Aurantivirga* PULs were present. The alginate PULs contained varying combinations of PL6, PL7 and PL17 alginate lyase genes (Fig. 4.4K-L). Such a PUL has been described in *G. forsetii* KT0803^{T}, and is also found in numerous other marine *Flavobacteriia* (e.g. Kappelmann et al. (2018)).

4.4.4 Mash-cluster PUL repertoires and abundance patterns

PUL numbers per individual Mash-cluster ranged from zero to eight with the applied filtering criteria (Fig. 4.5). Mash clusters with the broadest PUL spectra belonged to the *Aurantivirga*, *Formosa* and *Polaribacter* clades. All other Mash-clusters contained at most three PULs, most of which were targeting α- or β-glucans. This separation of clades into narrower vs. broader PUL spectra was also reflected in the individual clade's abundance patterns.

The spring bloom *Bacteroidetes* community at Helgoland showed a clear shift from a relatively stable low diversity pre-bloom community with few PULs towards a more flexible, seemingly stochastic community with more diverse PULs during mid-blooms. In all three years the pre-bloom community was dominated by clades with streamlined PUL repertoires. Mash-clusters reaching at least ≥1 RPKM at pre-bloom time-points carried between zero to three PULs, the majority of which targeted α- or β-glucans (Fig. 4.6A). Mash-clusters 12 (*Cryomorphaceae*), 14 (MAG-121220-bin8) and 15 (unclassified *Flavobacteriaceae*) were among the *Bacteroidetes* with the highest pre-bloom abundances, reaching up to 17 RPKM (∼8.5% relative abundance). Mash-clusters 12 and 14 carried α-glucan PULs and Mash-clusters 12 and 15 carried β-glucan PULs

(Fig. 4.5,Fig. 4.6). Overall, PUL repertoires at pre-bloom time-points indicated that the pre-bloom *Bacteroidetes* community predominantly targeted rather simple glycans.

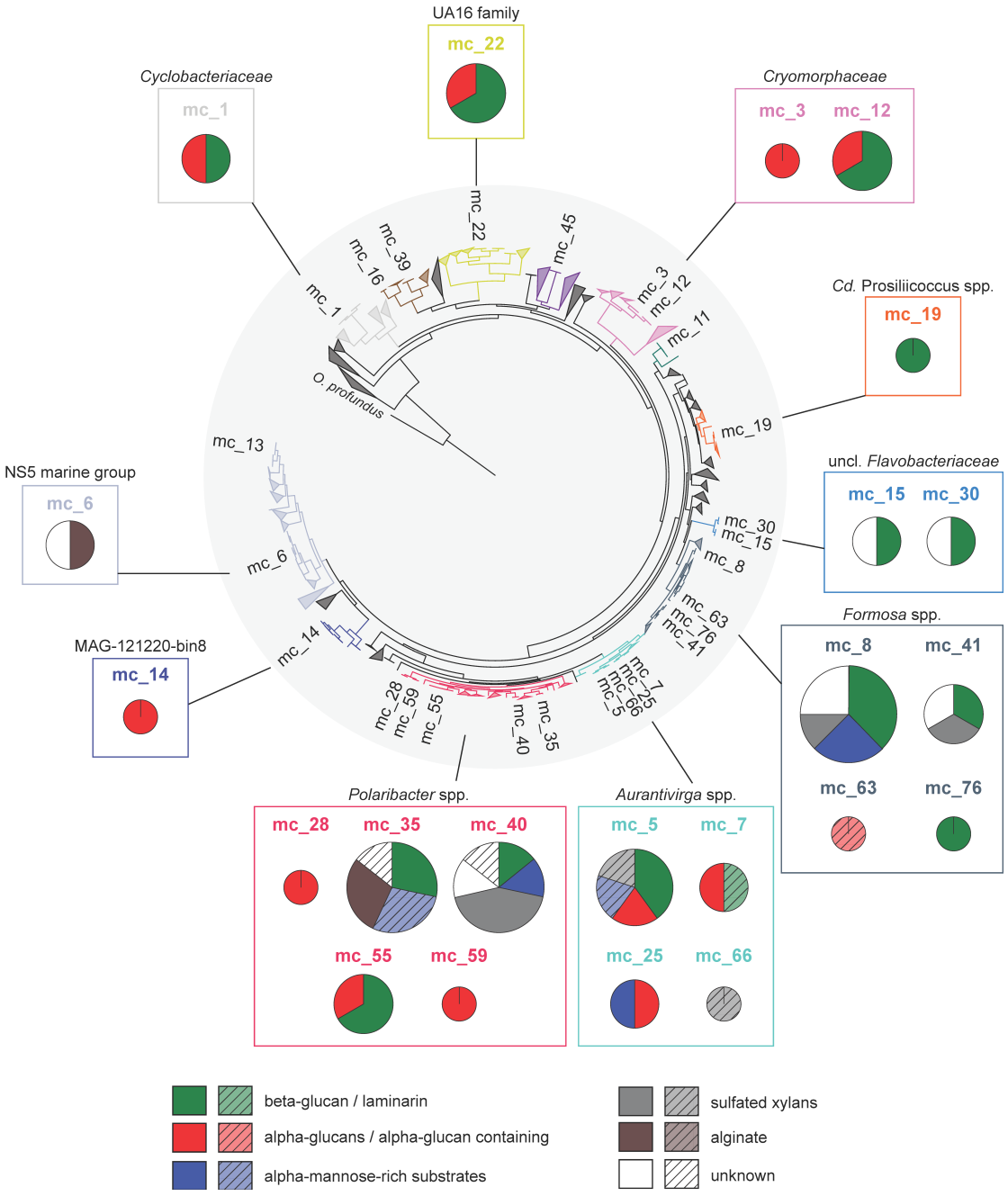


Figure 4.5: PUL repertoires of abundant Mash-clusters. Pie charts depict the predicted PUL substrates for each Mash-cluster and are sized according to PUL numbers (minimum: 1; maximum: 8 in mc_8). Mash-clusters that lack predicted PULs are not visualized. Box colors indicate PULs consistently binned in the same Mash-cluster, while hatched colors indicate putative PULs present in respective Mash-clusters (majority, but less than 50% of the same SusC/D binned into Mash-cluster).

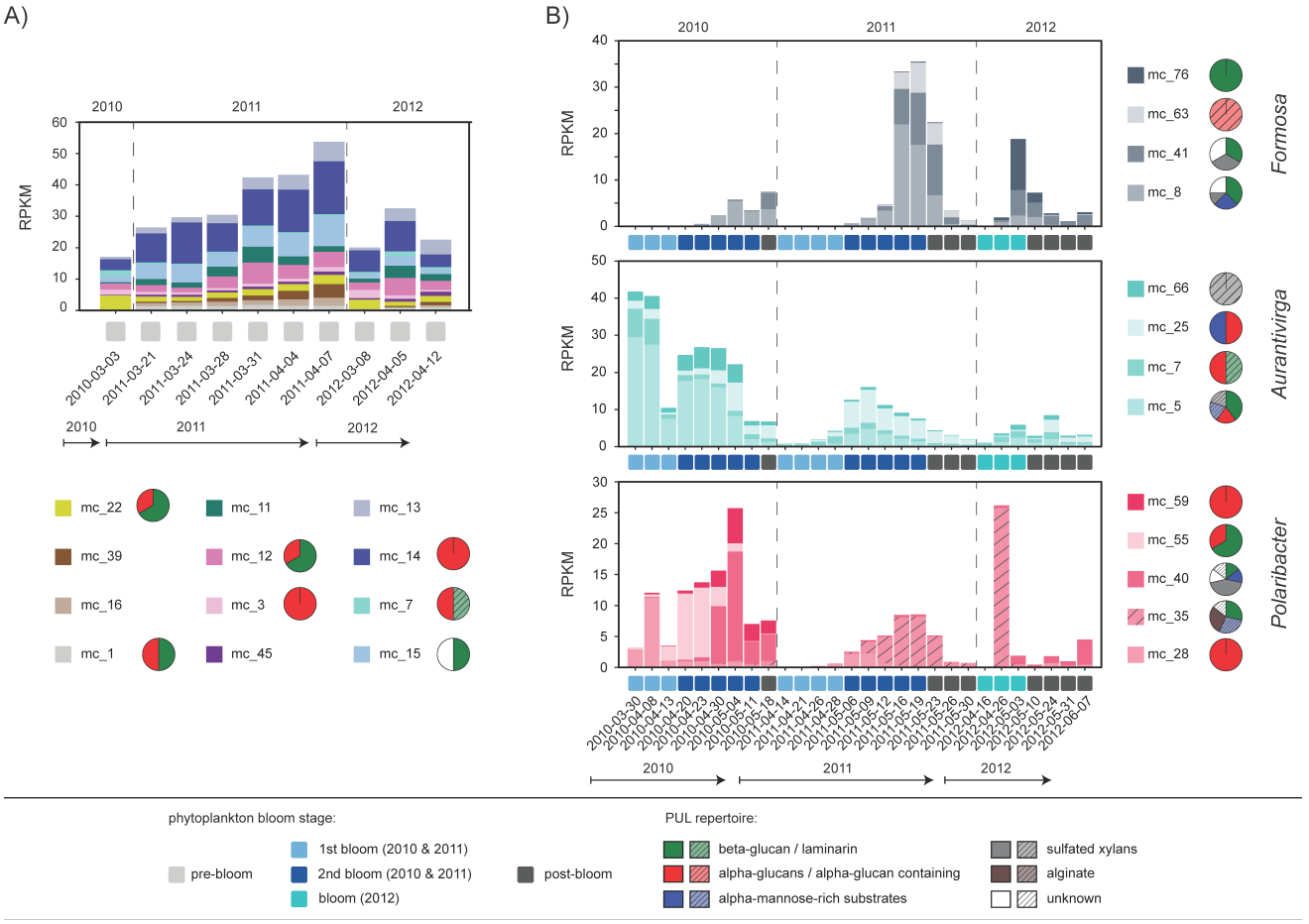


Figure 4.6: Abundance patterns of Mash-clusters at (A) pre-bloom and (B) (mid-) bloom and post-bloom phytoplankton bloom stages. Mash-cluster abundances are shown as reads-per-kilobase per million (RPKM) values in all bar charts. (A) includes all Mash-clusters with ≥1 RPKM at at least one pre-bloom time-point and (B) the abundant *Formosa*, *Polaribacter* and *Aurantivirga* Mash-clusters. Pie-charts depict predicted PUL repertoires of respective Mash-clusters. Abundance patterns of other abundant Mash-clusters are shown in Supplementary Figure 4.10.

The mid-bloom bacterioplankton response was more variable in all three years, though often species of the same clades reached high relative abundances. The PUL repertoires of mid-bloom *Bacteroidetes* Mash-clusters were broader compared to those that dominated pre-bloom phases. In the initial mid-bloom phase of 2010 we detected few highly abundant species. Mash-clusters 5 (*Aurantivirga*), 12 (*Cryomorphaceae*), 15 (unclassified *Flavobacteriaceae*), 19 (*Cd.* Prosiliicoccus) and 28 (*Polaribacter*) reached abundances well above 10 RPKM (∼5% relative abundance; Fig. 4.6, Fig. 4.10). All of these Mash-clusters carried PULs, targeting either α-glucans (mc_5, mc_12, mc_28) or laminarin/β-glucans (mc_5, mc_12, mc_15, mc_19). Additionally Mash-cluster 5 putatively targets mannose-rich substrates and sulfated xylans. These Mash-clusters, except for mc_5, leveled off in relative abundance towards the next stage of the 2010 bloom, where Mash-cluster 55 (*Polaribacter*) reached abundances of about 12 RPKM. Mash-cluster 55 carried PULs for both β-glucans and α-glucans. Towards the end of the bloom at the last three sampling time-points Mash-clusters 8 (*Formosa*) and 40 (*Polaribacter*) became abundant. Both showed similar PUL repertoires, as both encoded PULs targeting β-glucans, mannose-rich substrates, sulfated xylans as well as unknown substrates.

The 2011 phytoplankton bloom was less intense than in 2010 and so was the bacterioplankton's response (Teeling et al., 2016). For example, the response of *Cd.* Prosiliicoccus (mc_19), *Polaribacter* and *Aurantivirga* was less intense than in 2011 (Fig. 4.6, Fig. 4.10). Furthermore, the species composition of the *Polaribacter* and *Aurantivirga* clades were different. The *Aurantivirga* clade reached a maximum of 20 RPKM (∼10% relative abundance) with Mash-clusters 5 and 25 contributing about equally. The *Polaribacter* clade reached only up to 10 RPKM and was dominated by Mash-cluster 35, which had a broad PUL repertoire targeting β-glucans, putative mannose-rich substrates and alginates (Fig. 4.5, Fig. 4.6). In contrast, *Formosa* species dominated the *Bacteroidetes* community and reached up to 35 RPKM towards the end of the secondary bloom. Mash-clusters 8, 41 and 63 contributed to the abundances, although they showed different PUL repertoires (Fig. 4.5, Fig. 4.6).

In 2012 both the overall phytoplankton bloom and the *Bacteroidetes* response were even less intense (Teeling et al., 2016). Only Mash cluster 35 (*Polaribacter*) and the *Formosa* clade reached RKPM values above ten at only two time points (Fig. 4.6). There, *Formosa* was dominated by Mash-clusters 41 and 76, both of which carried β-glucan PULs with mc_41 additionally targeting sulfated xylans and as yet unknown substrates.

Overall *Aurantivirga*, *Cd.* Prosiliicoccus, *Formosa* and *Polaribacter* Mash-clusters were absent, or present at only low abundances (*Aurantivirga*) during the pre-bloom time-points but reached high abundances during the blooms. Thus these clades constitute

major bloom responders (Teeling et al., 2016). With the exception of *Cd.* Prosiliicoccus (Francis et al., 2018) these clades featured broad PUL repertoires. *Formosa* and *Aurantivirga* harbored PULs for four of the major substrates targeted by *Bacteroidetes* during phytoplankton spring blooms and *Polaribacter* even all five including alginate.

4.5 Discussion

Marine *Bacteroidetes* have been regularly observed as major responders during phytoplankton blooms and usually outcompete other bacterial clades during early bloom phases (e.g. Needham et al. (2018), Needham & Fuhrman (2016), Taylor et al. (2014)). The ability to decompose high-molecular-weight (HMW) organic matter (Kirchman, 2002) using PULs with efficient SusC/D-like uptake systems is likely pivotal for the competitiveness of many of these *Bacteroidetes*. Plenty of PULs have been detected in isolated *Bacteroidetes* strains (e.g. Kabisch et al. (2014), Kappelmann et al. (2018), Tang et al. (2017), Valdehuesa et al. (2018)). However, many of these are associated with macro-algae (e.g. Barbeyron et al. (2016b)), whereas only a few are representatives of abundant free-living *Bacteroidetes* during micro-algae blooms (Kappelmann et al., 2018, Unfried et al., 2018, Xing et al., 2015). Therefore their PUL repertoires do not indicate, which PULs predominate during such blooms.

Recent improvements in metagenome sequencing, assembly and binning have enabled large-scale retrieval of PULs from natural bacterioplankton and subsequent linkage to distinct species. This approach works particularly well for bacterioplankton during early spring bloom stages, where there is low evenness due to stark proliferation of few well adapted species, some of which are almost clonal (e.g. Avci et al. (2017)). Very much on the contrary, marine sediments represent much more demanding habitats due to much higher species richness and evenness. Metagenomes obtained from sandy sediments at the same North Sea sampling site, for example, did almost not assemble (unpublished data).

Our analysis of a dense time series of deeply sequenced bacterioplankton metagenomes in combination with state-of-the-art binning and subsequent PUL prediction and annotation circumvents the isolation problem. However, it also entails a certain margin of error, as PUL functions are predicted based on similarity searches against reference databases and not on laboratory-based experiments with dedicated bacterial strains. Likewise the sheer size of the dataset (38 metagenomes with a total of 9.9 Gbp) required a rather rigid automated PUL prediction, which neglected non-canonical PULs devoid of a *susC/D* pair. It has been shown that some PULs have *susC/D* pairs that are separated from the corresponding CAZymes elsewhere in the genome, (e.g. Ficko-Blean

et al. (2017)). Likewise PULs from sparsely occurring Mash-clusters were excluded as we restricted our analysis to PULs that occurred in at least four metagenomes. Still the Mash-clusters we describe provide first time insights into the most prevalent PULs and their predicted algal polysaccharide targets within the largely uncultivated planktonic *Bacteroidetes* community during spring phytoplankton blooms in the Southern North Sea.

The analyzed *Bacteroidetes* communities were dominated by few clades, such as *Cd.* Prosiliicoccus, *Formosa*, *Polaribacter* and *Aurantivirga*. Mash-cluster analysis demonstrated recurrence of these clades in 2010 to 2012 with some inter-annual variability, substantiating an initial analysis of a much smaller subset of the metagenomes (Teeling et al., 2016). This recurrence suggests a high level of specialization on bloom events. Rather small genome sizes with limited numbers of PULs enable these *Bacteroidetes* to quickly respond to phytoplankton blooms with fast growth rates, while targeting only specific subsets of the glycans that algae produce. The average genome size of our Mash cluster representatives was about 1.5 Mbp lower compared to genome sizes of isolated North Sea strains of 3.83 Mbp (Kappelmann et al., 2018). Two published single cell genomes of the NS5 and NS3a marine groups (*Flavobacteria bacterium* MS024-2A; *Flavobacteria bacterium* MS024-3C) are closely related to the mc_13 and mc_11 in our study. These two species also feature small genomes and narrow ecological niches (Woyke et al., 2009). It has been speculated that this might be the reason, why they resist cultivation, even though they are ubiquitously abundant (Woyke et al., 2009). Considering, that genome streamlining entails a reduction of physiological flexibility (Swan et al., 2013), it is not unexpected that a lot of free-living marine *Bacteroidetes* resist conventional isolation techniques.

The majority of abundant PULs that we describe constitute a subset of the PUL spectrum that has so far been described in isolated *Bacteroidetes* strains (Kappelmann et al., 2018, Unfried et al., 2018, Xing et al., 2015). These PULs are limited to five major substrate classes of which β-glucans/laminarin and α-glucan-containing substrates are both most abundant as well as present throughout all bloom periods. PULs targeting these simple glycans represent more than half of the described metagenomic PULs (50/131 β-glucans/laminarin; 22/131 α-glucan-containing substrates), which is substantially more than compared to only about 25% of PULs in the genomes of isolated North Sea Bacteroidetes strains (Kappelmann et al., 2018)). This suggests that both β-glucans/laminarin and α-glucan-containing substrates make up the majority of substrates available to the bacterial community during phytoplankton blooms. Laminarin is the major storage compound of diatoms and so are slightly different types of β-glucans in other phytoplankters (Myklestad & Granum, 2009). Alpha-glucans on the other hand act as storage compound in bacterial and animal cells (e.g. glycogen). Thus

laminarin and α-glucans are constantly processed and released during bloom events due to factors such as grazing, viral lysis or autologous cell death that influence microbial mortality (reviewed in Brum et al. (2014)).

We observed an average of only 2.2 PULs per Mash-cluster compared to an average of 7.5 PULs for genomes of isolated North Sea *Bacteroidetes* strains (Kappelmann et al., 2018), and similar numbers in other marine *Bacteroidetes* (e.g. Barbeyron et al. (2016b)). Notable exceptions were the *Polaribacter*, *Formosa* and *Aurantivirga* Mash-clusters, with some species exhibiting the highest PUL numbers (maximum: eight) and most diverse predicted substrate spectra. These clades all share the potential to degrade α-mannose-rich substrates and sulfated xylans/xylose-rich substrates. Co-eluting sugars mannose and xylose have been detected as second most abundant monosaccharides in the total combined carbohydrates during the 2010 bloom (Sperling et al., 2017). Still the glycan niches of the most abundant bacteroidetal bacterioplankters are rather narrow, which is why the remineralization of algal glycans is a concerted effort of many of these clades.

Throughout the entire time-series some clades featured rather constant PUL repertoires, e.g. *Cyclobacteriaceae*, whereas there was considerable compositional change in others, e.g. within the broad *Polaribacter* clade. We observed a clear dominance of PULs targeting simple glycans (β-glucans/laminarin and α-glucan-containing substrates) in pre-bloom communities, whereas amidst blooms also more complex polysaccharides were targeted. We hypothesize that this is the result of two effects. First, bacteria will on overall prefer easily degradable substrates such as simple storage glycans over biochemically more demanding ones. Second, the availability of more complex polysaccharides increases over a blooms' course due to increasing algae mortality rates.

It is still an open question, which selective effects favor one of two species with similar PUL repertoires over the other. Species with similar PUL repertoires might still prefer different polysaccharides. For example, co-cultivation of two gut *Bacteroidetes* suggested that their glycan preferences are genetically hard-wired (Tuncil et al., 2017). In a similar fashion, slight subtle differences in polysaccharide composition might also be a contributing factor. High-resolution PUL in situ expression data over the entire course of a phytoplankton bloom will be necessary to further enhance our understanding, as abundance does not necessarily equate to high activity (e.g. Bryson et al. (2017)).

The composition dynamics of phyto- and bacterioplankton communities during blooms are complex, as is the resulting glycan turnover biochemistry. Nonetheless, a limited number of bacterioplankton clades prevail during bloom conditions that carry a limited number of abundant PULs, which in term target a limited number of major glycan substrates. This means that to attain a fundamental understanding of the bulk of glycan-mediated carbon flow during phytoplankton bloom events is within reach.

4.6 Acknowledgements

We thank Sabine Kühn for DNA extractions and Ivaylo Kostadinov of GFBio (http://www.gfbio.org) for sequence data deposition support. Genome sequencing and assembly was conducted in the framework of the Community Sequencing Project COGITO (CSP 998) by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, and is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. This study was funded by the Max Planck Society and supported by the Deutsche Forschungsgemeinschaft (DFG) in the framework of the research unit FOR2406 'Proteogenomics of Marine Polysaccharide Utilization (POMPU)' by grants of Rudolf Amann (AM 73/9-1), Hanno Teeling (TE 813/2-1), and Thomas Schweder (SCHW 595/10-1). Karen Krüger and T. Ben Francis are members of the International Max Planck Research School of Marine Microbiology (MarMic).

Supplementary material

4.7 Supplementary text

4.7.1 Supplementary materials and methods

4.7.1.1 Metagenome sequencing, assembly and automated binning

38 surface seawater metagenome samples were sequenced at the Department of Energy Joint Genome Institute (DOE-JGI, Walnut Creek, CA, USA) as previously described for a subset of ten of these (Teeling et al., 2016) (sampling dates: 2010/03/03; 2010/04/08; 2010/05/04; 2010/05/18; 2011/03/24; 2011/04/28; 2011/05/26; 2012/03/08; 2012/04/16; 2012/05/10). Additional 28 samples were sequenced (sampling dates: 2010/03/30; 2010/04/13; 2010/04/20; 2010/04/23; 2010/04/30; 2010/05/11; 2011/03/21; 2011/03/28; 2011/03/31; 2011/04/04; 2011/04/07; 2011/04/14; 2011/04/21; 2011/04/26; 2011/05/06; 2011/05/09; 2011/05/12; 2011/05/16; 2011/05/19; 2011/05/23; 2011/05/30; 2012/04/05; 2012/04/12; 2012/04/26; 2012/05/03; 2012/05/24; 2012/05/31; 2012/06/07). Latter samples were sequenced on the Illumina HiSeq 2500 platform (Illumina, San Diego, CA, USA) with four samples pooled per flow-cell lane, resulting in about one fourth of reads per sample compared to the initial samples (see Table S1 - online for details).

Quality filtering and trimming of raw reads, metagenome assembly and binning using CONCOCT (Alneberg et al., 2014) was performed as described previously (Francis et al., 2018). In brief, BBDuk v.35.14 (http://bbtools.jgi.doe.gov) was used to remove TruSeq adapters and low quality reads (options: ktrim=r, k=28, mink=12, hdist=1, tbo=t, tpe=t, qtrim=rl, trimq=20, minlength=100). Metagenomic datasets were assembled individually using metaSPAdes v3.10.0 (Nurk et al., 2017) with kmer lengths of 21, 33, 55, 77, 99 and read-error correction enabled. Contigs below 2.5 kbp were excluded from further analyses. Separate binning of contigs from each assembly was performed using CONCOCT (Alneberg et al., 2014) integrated in the anvi'o metagenomic workflow (Eren et al., 2015). Read coverage profiles were generated by mapping SPAdes error-corrected

reads of the respective sampling date and four additional SPAdes error-corrected read sets from metagenomic samples of the same year to contigs ≥2.5 kbp (Table S1 - online). BBMap v35.14 (http://bbtools.jgi.doe.gov) was used for read mapping in fast mode, with a minimum mapping identity (minid) of 0.99 and identity filter (idfilter) of 0.97. Subsequently sequence alignment map files were converted to binary format, filtered, sorted and indexed using SAMtools v1.2 (Li et al., 2009). Finally, sequence alignment map files were filtered: unmapped reads (-F 4), reads mapped with low quality <10 (-q) and reads found to be PCR duplicates (VALIDATION_STRINGENCY=LENIENT) were removed using Picard tools v1.133 (http://broadinstitute.github.io/picard).

4.7.1.2 Metaproteome SusC and SusD extraction

Metaproteome SusCs and SusDs were extracted from the metaproteome data by comparing all identified expressed protein sequences (total: 23 917) to the Hidden Markov Models of TIGRFAM profile (TIGR04056) for SusC/RagA and Pfam profiles for SusD (PF07980, PF12741, PF14322, PF12771). Hits were post-filtered using the hmmscan-parser script with the exception of the filtering step that excludes hits with less than 30% coverage of the model. All identified proteins were then integrated into the SusC- and SusD-protein trees and considered as representing in situ expression, if they were on amino acid level >90% similar to the metagenomic *susCs* and *susDs*.

4.7.2 Supplementary results

4.7.2.1 PULs in *Bacteroidetes* Mash-clusters

Unknown substrates The SusC/D protein trees showed a cluster close to the laminarin/β-glucan cluster, whose corresponding PULs had a combination of CAZymes so far not described in literature. These PULs encode similar CAZyme families as laminarin PULs of variant two (3x GH17, GH30_1), but instead of a GH16 a GH92 gene is present. Other PULs in this cluster contained three GH92 genes and either a GH2 or two GH3 genes. A similar PUL has been predicted by PUL-DB for e.g *Flavobacterium johnsoniae* UW101, where instead of a third GH92, a GH125 gene is present that has exo-α-1,6-mannosidase activity as only known activity (Gregg et al., 2011). These PULs occurred in both the *Formosa* genus and Mash-clusters 15 and 30 (unclassified *Flavobacteriaceae*), but substrate specificity remains as yet unclear.

Mash-cluster 57 contained the longest PUL in the entire dataset. This PUL harbored two *susC/D* pairs (one with two *susC* homologues), and comprised 20 CAZyme and sulfatase genes. Mash-cluster 57 is a lowly abundant member of the NS5 marine group

and was binned from post-bloom time-points in 2011 and 2012 (Fig. 4.9). The PUL's CAZyme composition suggests a sulfated mixed-glucan containing rhamnose, galactose and mannose. The PUL encodes a GH92 (exo-α-mannosidase activity) and several glycoside hydrolase families with enzymes acting on rhamnose or rhamnose-galactose containing substrates, such as two GH105, a GH78 and a GH28.

4.8 Supplementary figures

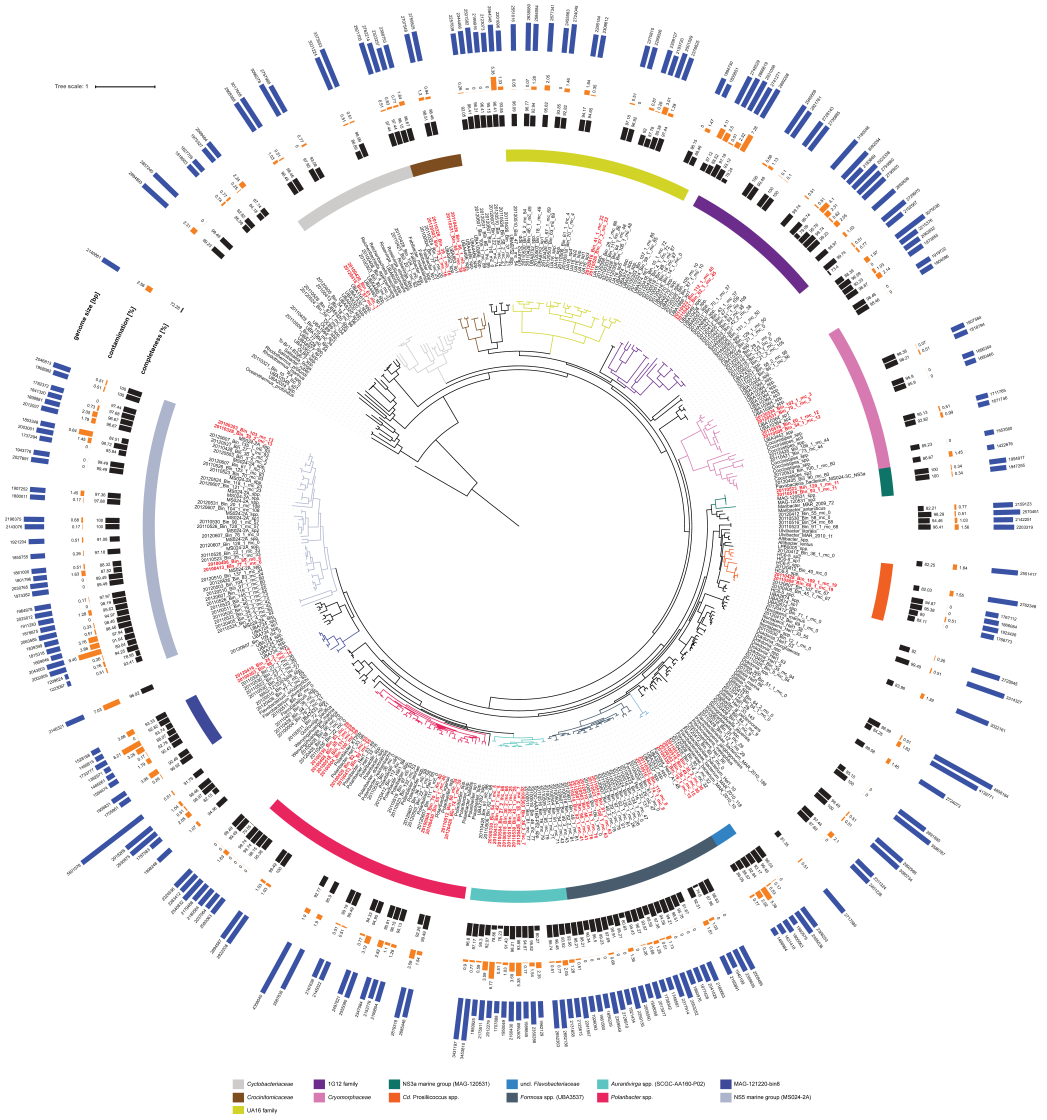


Figure 4.7: Detailed maximum-likelihood tree of Mash-clusters based on concatenated marker proteins according to the GTDB$_{tk}$ genome phylogeny (Parks et al., 2018). Mash-clusters (mc) were only included in the tree that contained at least two MAGs ≥70% completeness and are represented by two MAGs each. Additional 29 singleton MAGs with ≥70% completeness and ≤10% contamination were included (labeled as mc_0). Mash-clusters reaching high abundances (>5 RPKM at one time-point or 38 RPKM at all time-points combined) during the sampling period are highlighted together with their taxonomic affiliations. Black and orange colored bars depict completeness and contamination values of respective MAGs, while blue bars indicate MAG genome sizes. Scale bar: mean number of amino-acid substitutions per site. Outgroup: *Oceanithermus profundus*.

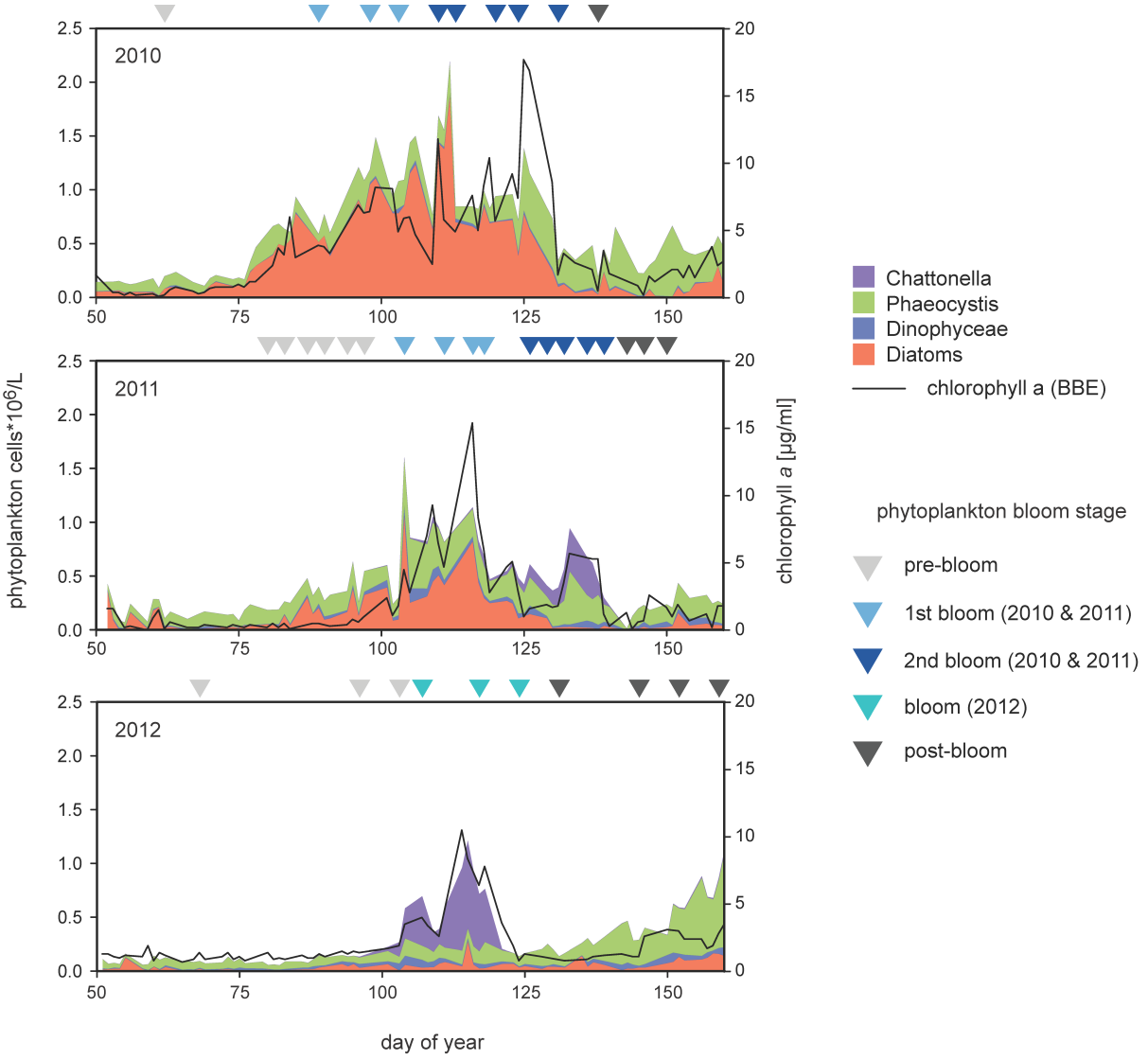
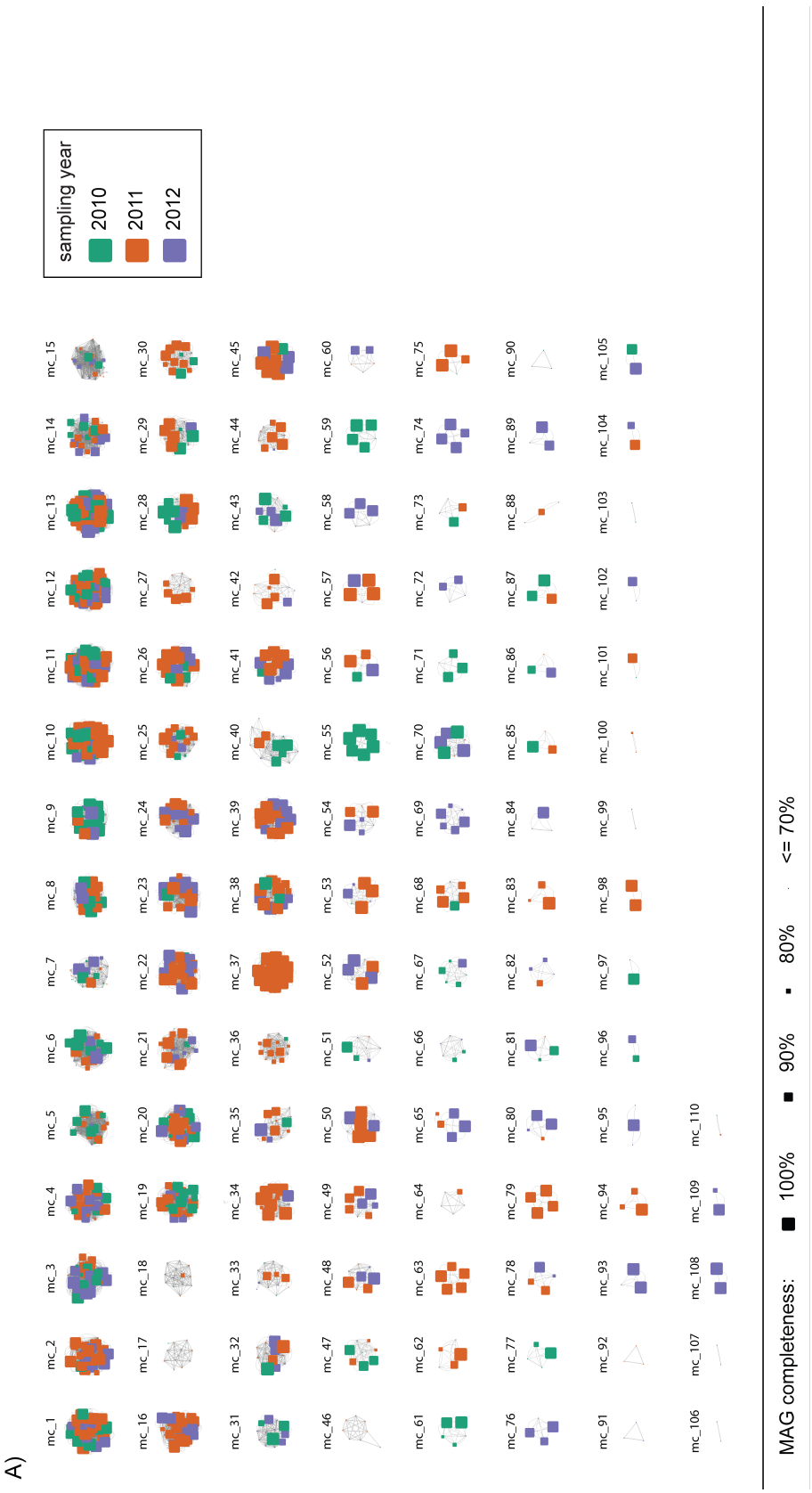


Figure 4.8: Phytoplankton blooms in the years 2010 to 2012 depicted both by chlorophyll *a* and cell counts of major phytoplankton groups. Triangles on top indicate metagenome sampling time-points and are color coded based on phytoplankton bloom stage.


A)
sampling year
2010
2011
2012
MAG completeness:
100%
90%
80%
<= 70%

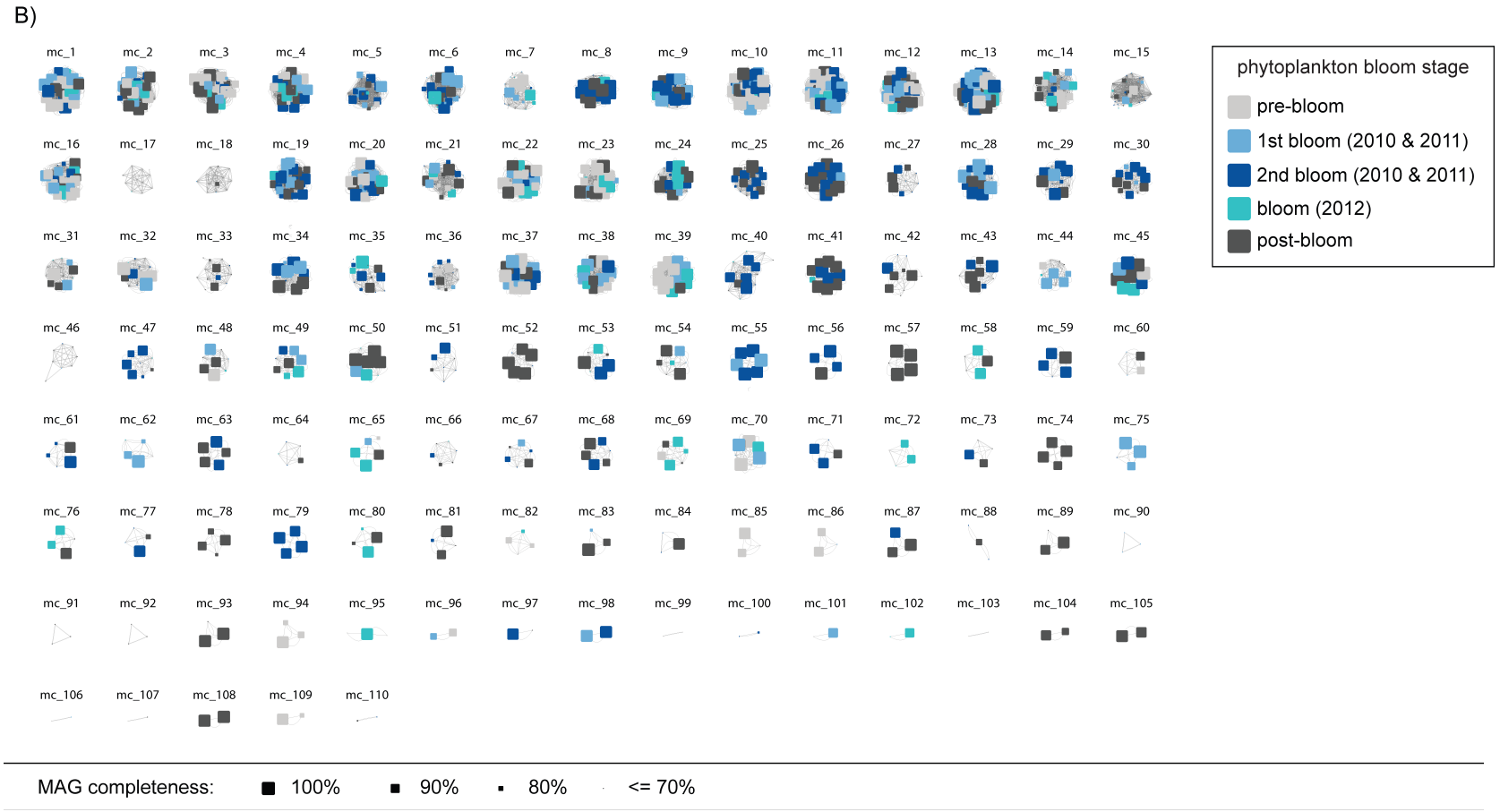


Figure 4.9: Composition of all Bacteroidetes Mash-clusters with respect to sampling year (A) and phytoplankton bloom stage (B). Squares represent individual MAGs with sizes corresponding to completeness, while gray lines indicate Mash distances $\leq$0.05, and thus the approximate species connections for MAGs in each Mash-cluster. (A) Color coding indicates the year in which individual MAGs were retrieved. (B) Color coding represents phytoplankton bloom stages.

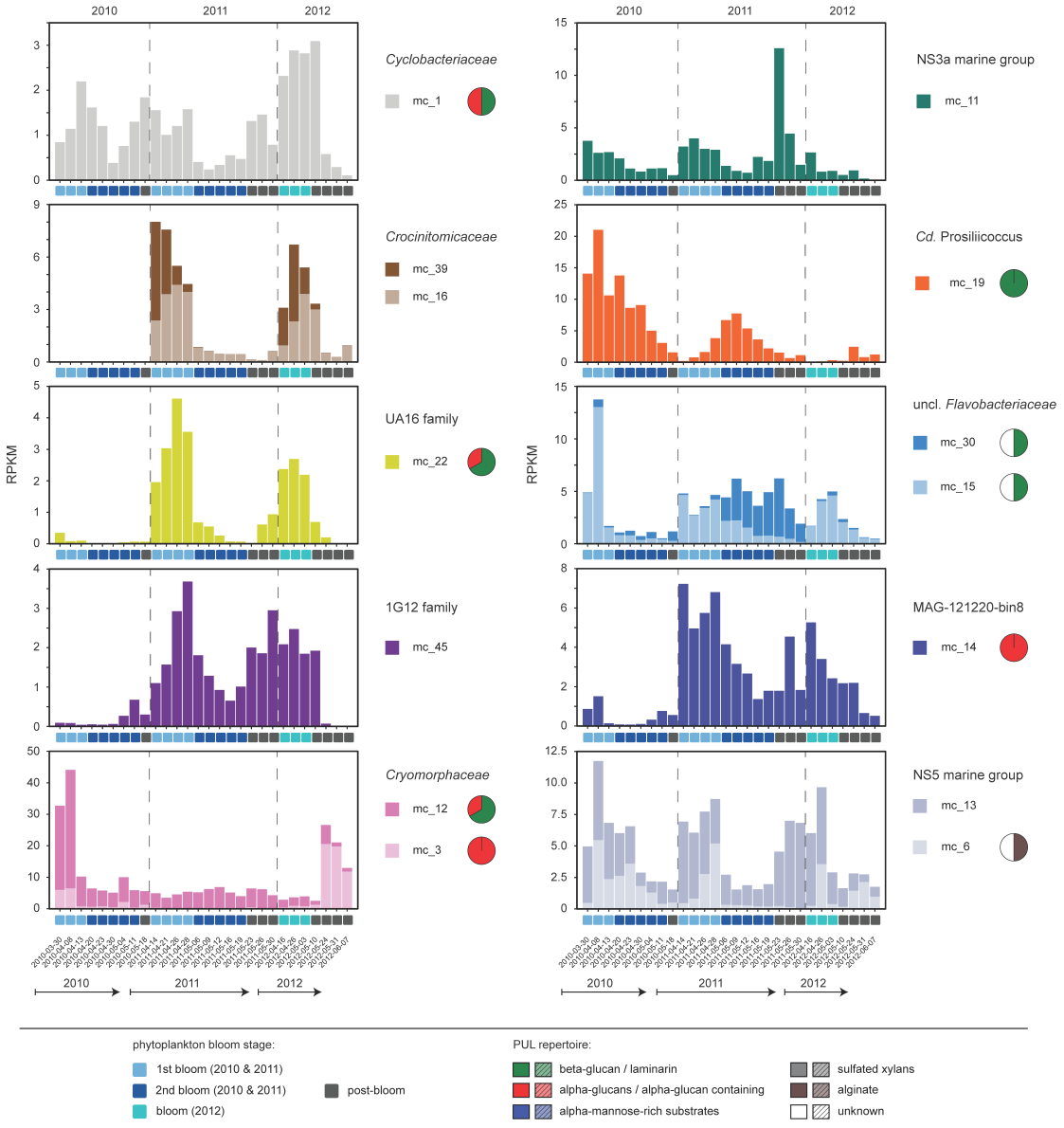


Figure 4.10: Abundance pattern of Mash-clusters at bloom and post-bloom phytoplankton bloom stages. Mash-cluster abundances are shown as reads-per-kilobase per million (RPKM) values in all barcharts. (Mid-) bloom and post-bloom abundances are shown for all abundant Mash-clusters from clades not presented in Figure 4.6. Pie-charts depict predicted PUL repertoires of respective Mash-clusters.

Chapter 5

Discussion and outlook

As of this writing polysaccharide degradation capabilities of PULs have been studied in marine *Bacteroidetes* isolates for about two decades. These studies though, focused on few model organisms only, whereas a more holistic overview about environmentally relevant PULs is as yet lacking. Individual species' capabilities to degrade polysaccharides, and thus their PUL repertoires, are distinct and can reflect a species' habitat as well as its ecological niche therein. The macroalgae-associated isolate *Zobellia galactanivorans* DsiJ^{T}, for example, harbours a total of 141 GHs, 15 PLs and 18 CEs and a total of 50 PULs, targeting diverse polysaccharide substrates (Barbeyron et al., 2016b). Free-living marine *Bacteroidetes* on the other hand, often comprise a limited set of PULs and degradative CAZymes (Barbeyron et al., 2016b, Xing et al., 2015). These differences in lifestyle are often manifested in smaller genome sizes of free-living *Bacteroidetes* and larger genome sizes in macroalgae- or particle-associated *Bacteroidetes* (Xing et al., 2015). Small genome sizes and a reduced physiological flexibility might be one reason why many of the free-living *Bacteroidetes* resist cultivation (Swan et al., 2013, Woyke et al., 2009), even though they are often among the most abundant organisms that respond to spring phytoplankton blooms. Thus it is of great importance to study these free-living organisms to shed light on the polysaccharide degradation capabilities of uncultured marine *Bacteroidetes*.

5.1 Recurrence of spring bloom responders at Helgoland Island

The investigation of the bacterioplankton's response to spring phytoplankton blooms at the island of Helgoland started with a pilot study in 2009 (Teeling et al., 2012). This

study revealed that bacteria responding to the diatom dominated bloom were characterized by a few prominent clades which showed a swift successional pattern. Based on functional traits analysed by metagenomics and metaproteomics it was suggested that the response of these abundant clades was predominantly "bottom-up" controlled by available substrates. These mostly algae-derived substrates were hypothesized to change over the course of the bloom, creating a series of ecological niches, in which different bacterial clades could thrive (Teeling et al., 2012). To further investigate the succession of heterotrophic bacteria equipped with the genetic toolkit for the usage of algal derived OM, and to shed light on the recurrence and consistency of this response pattern over several spring phytoplankton blooms, the study was extended for another three years. This extended study, presented in Chapter 2, covered four consecutive years (2009-2012) of spring phytoplankton blooms at the island of Helgoland. Even though these four investigated years differed in phytoplankton species composition and bloom intensities, the response of the bacterioplankton community was remarkably consistent in all four years with most of the same clades recurrently present. Among the recurrent *Bacteroidetes* clades were *Formosa*, *Polaribacter*, the NS3a and the NS5 marine groups, *Tenacibaculum*, *Ulvibacter*, and the *Cryomorphaceae* VIS6 clade (Chapter 2, Fig. 2.3 and Fig. 2.4). Despite variations in appearance time and maximum abundance, these clades together made up the majority of the *Bacteroidetes* community, and most of these clades recurred in all four years.

Many of the detected *Bacteroidetes* clades are commonly found related to phytoplankton blooms in other parts of the world, such as the *Cryomorphaceae* and the NS3a and NS5 marine groups, which have been detected during dinoflagellate blooms (Tan et al., 2015, Yang et al., 2015). Similarly, the NS5 marine group, the *Polaribacter*, *Ulvibacter* and *Formosa* clades were detected after a diatom bloom offshore of Santa Catalina Island, California (Needham et al., 2018, Needham & Fuhrman, 2016). Amplicons with >99% identity to the 16S rRNA of an *Ulvibacter* species, now reclassified as *Candidatus* Prosiliicoccus vernus (Appendix C), could be detected globally and were more often than not related to phytoplankton bloom situations (Appendix C, Fig. 3.4). While some of these clades are also found in abundance during other times of the year (e.g. Appendix B and C), some individual sub-OTU level oligotypes were shown to be related to spring blooms by network analysis (Chafee et al., 2017). Thus these clades most likely represent species that are highly specialized on the usage of phytoplankton-derived OM. Functional potentials of these major *Bacteroidetes* clades could only be addressed at genus-level in Chapter 2. Nevertheless, we found that the overall functional potentials of the *Flavobacteriia* (a class of the *Bacteroidetes*) showed an almost fingerprint-like CAZyme repertoire throughout the course of the bloom (Chapter 2, Fig. 2.6), which was unaffected by the distinct *Bacteroidetes* clades dominating the community at the

individual sampling time points. This suggests that the *Bacteroidetes* community as a whole is uniform in their polysaccharide degradation potential and that different *Bacteroidetes* species might substitute each other throughout the course of the bloom.

The extended metagenomic dataset that I analysed in Chapter 4 allowed an increase in resolution regarding sampling and taxonomy. By adding another 28 metagenomes for the years 2010 to 2012, I was able to focus on weekly rather than monthly shifts in the *Bacteroidetes* community. In addition, the taxonomic resolution increased down to species level as I was able to extract MAGs that represented about 120 *Bacteroidetes* species. 75% of these 120 species and about 90% of the 27 abundant species belonged to clades already detected in Chapter 2. Many of these MAGs were detected in multi-year Mash-clusters (approximate species clusters; Fig.4.2), which were often constantly present during the phytoplankton blooms, and which support the recurrence described in Chapter 2. Exceptions here were primarily some of the Mash-clusters from the *Aurantivirga* (previously part of the *Polaribacter* clade), *Formosa* and *Polaribacter* genera (Chapter 4, Fig. 4.2) that could only be detected in a single year, and were replaced by other single- or multi-year Mash-clusters of the same genus in other years. This higher species variation in *Formosa* and *Polaribacter* would support the more variable CAZyme repertoires of these clades detected in Chapter 2 (Fig. 2.6). *Polaribacter*, for example, showed the broadest CAZyme repertoire, and this repertoire was the most variable between different sampling dates. Thus the detected recurrence described in Chapter 2 and corroborated in Chapter 4 is a recurrence on genus-level, with a few highly adapted clades. These recurrently thrive after phytoplankton blooms, with some clades displaying a high diversity with regards to species composition.

Overall the recurrence of only a few highly adapted clades, and of 27 abundant species within these, corroborates that the community diversity in the 0.2-3 μm size fraction of bacterioplankton is constrained during springtime, as previously suggested by 16S rRNA analyses (Chafee et al., 2017). Consequently, the MAGs described in Chapter 4 likely cover the majority of the abundant *Bacteroidetes* species at our sampling site and allow us to describe the most prominent *Bacteroidetes* species and their functional repertoires. The fact that these clades are also globally detected in relation to phytoplankton blooms, suggests that the *Bacteroidetes* community diversity, adapted for the degradation of phytoplankton-derived substrates, might also globally be constrained to a few prominent clades, which would be interesting to investigate in more detail.

5.2 Automatic predictions of PULs

Advances in assembly algorithms, including the computational requirements, have allowed us to obtain improved metagenome assemblies during the time frame of my PhD. While the metagenome assemblies in Chapter 2 allowed us to study overall CAZyme repertoires, the study of PULs in these assemblies was more difficult and restricted to only sufficiently long contigs. Metagenomes from 2010 to 2012 in Chapter 2 were assembled using SOAPdenovo v.1 (Luo et al., 2012) with six different individually computed k-mer sizes. Contigs were then combined from these six separate assemblies (see Chapter 2 for details). Re-assemblies of the same datasets using metaSPAdes with iterative k-mer sizes yielded on average about twice as many assembled megabases for contigs larger than 500 bp (Fig. 5.1). Additionally, contigs were on average longer, which is in agreement with metagenome assembly tests performed using metaSPAdes and SOAPdenovo v.2 (Vollmers et al., 2017). Thus these metaSPAdes assemblies were better suited for PUL predictions as applied in Chapter 4, with longer contigs that allow for a much wider analysis of gene clusters than the on average shorter contigs in Chapter 2. In contrast to the positive effects of the increase of assembled data, this also necessitates the rather rigid filtering criteria applied in Chapter 4. This was first of all the focus on PULs encoding at least one SusC or SusD. This is a more conservative definition of a PUL, as some of the *Bacteroidetes* PULs have been described with a separate *susC/D* gene pair elsewhere in the genome (Ficko-Blean et al., 2017). Second was the filtering criteria based on the presence of the *susC* or *susD* genes in more than 10% of the 38 metagenomic datasets. Thus, the data regarding PULs presented in Chapter 4 provides a look at the most prevalent PULs present in the *Bacteroidetes* community.

The sheer amount of metagenome data generated in Chapter 2 and Chapter 4 requires automatic predictions of functions of interest, such as the genes involved in the degradation of polysaccharides, as it was applied in Chapter 4. Of course these automatic predictions cannot rival laboratory-based experiments with dedicated bacterial strains, but they have the advantage of giving a broad overview of uncultured organisms' functions that could then be tested in isolate strains carrying similar PULs. Even though PUL functions are predicted based on similarity searches against reference databases, these databases, such as the CAZy database (Lombard et al., 2014), contain functionally well described and manually annotated sequences. I applied a similar approach to PUL prediction in Chapter 4 as is also used by the automatic prediction of PULs in the PUL-DB (Terrapon et al., 2018, 2015), a database for experimentally verified and predicted PULs that is associated with the CAZy database. These automatic PUL predictions for isolate genomes were shown to be congruent with manually predicted PULs in respective isolates (Terrapon et al., 2015). Additional problems that arise using metagenomics for

PUL predictions is the limitation of the length of contigs, as PULs might be truncated at contig ends. Nevertheless, automatic PUL prediction in isolate genomes and especially metagenomes is a step towards a more holistic understanding of the relevant PULs and their targeted polysaccharide substrates present during spring phytoplankton blooms.

5.3 Polysaccharide utilization loci in marine *Bacteroidetes* - from isolates to metagenomes

In Chapter 3 we investigated the PUL repertoires of 53 *Bacteroidetes* isolates from the North Sea and were able to shed light on the overall PUL cosmos present in these North Sea isolates. The *Bacteroidetes* isolates were retrieved from different sources and represent a high genomic and phylogenetic diversity, from free-living to micro- and macro-algae associated species. Numbers of PULs ranged from 0 to 40, with an average of 7.5 PULs per genome. The isolate with the highest number of PULs and degradative CAZymes from the GH, PL and CE families was *Zobellia amurskyensis* MAR_2009_138, which is comparable to the PUL and CAZyme repertoires of *Zobellia galactanivorans* DsiJ^{T} (Barbeyron et al., 2016b). Overall the targeted polysaccharides or polysaccharide classes of 259 of the 400 detected PULs could be predicted based on the CAZymes and sulphatases present. In total these were 18 polysaccharide classes that included laminarin or similar β-glucans, α-glucans, and alginate, which were frequently distributed among the isolates. PULs targeting these three substrates have also been detected in other isolate studies (Bauer et al., 2006, Kabisch et al., 2014, Mann et al., 2013, Xing et al., 2015), including some gammaproteobacterial isolates (Neumann et al., 2015). PULs targeting polysaccharides such as pectin, chitin or rhamnose-, mannose-, fucose-, and xylose-rich substrates were less frequently detected in the 53 *Bacteroidetes* isolates. Overall the PUL repertoires were not uniform within genera and showed considerable intra-genus and inter-genus variation. This indicates that *Bacteroidetes* species maintain distinct glycan niches independent of their phylogenetic relationship. PULs are known to be exchanged through horizontal gene transfer (e.g. Hehemann et al. (2017), Naumoff & Dedysh (2012), Thomas et al. (2012)) and such events could be a factor for acquiring PULs that are more randomly distributed among the isolates, such as those targeting mannose- or xylose-rich substrates.

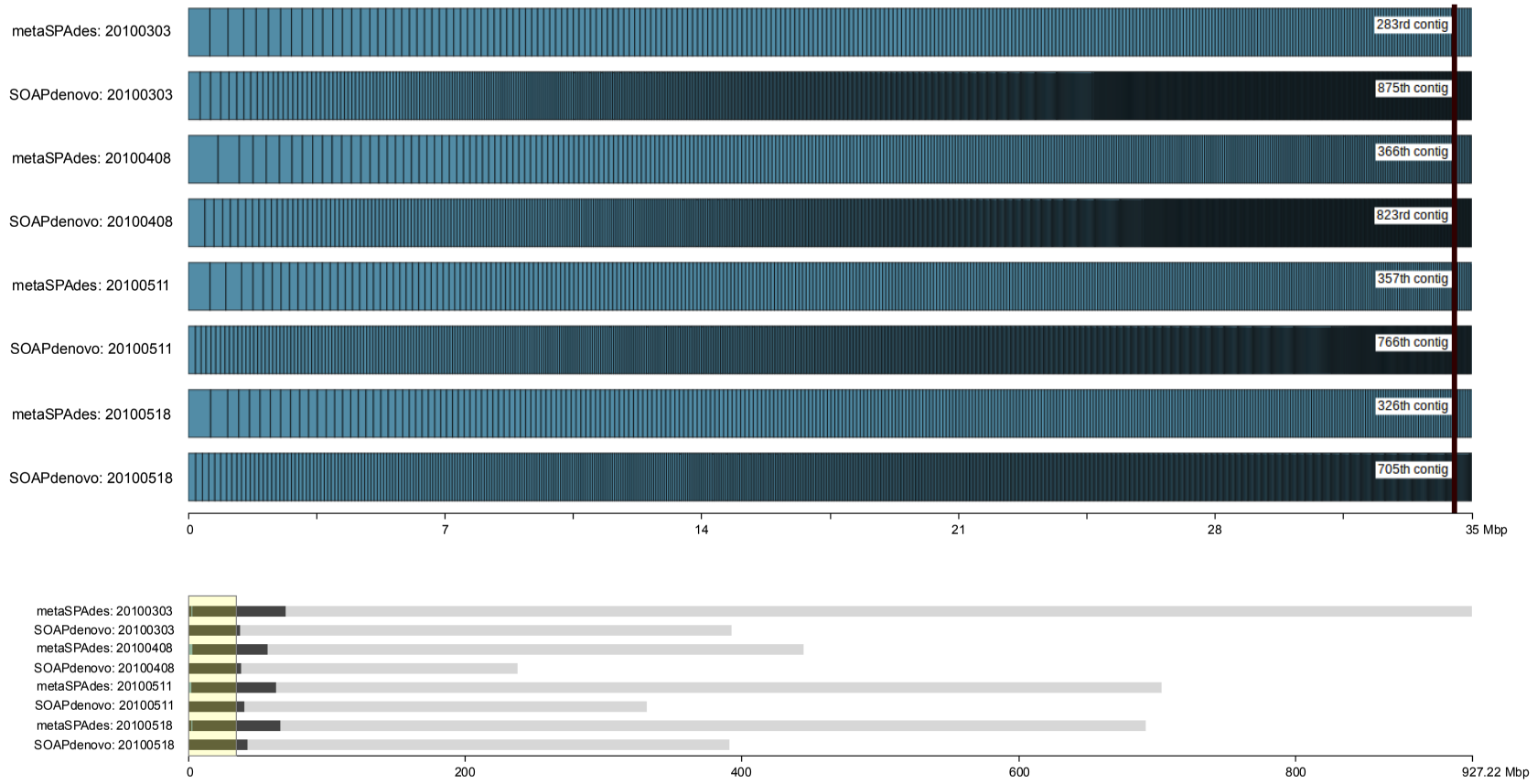


Figure 5.1: Comparison of selected metagenome assemblies from Chapter 2 and Chapter 4, based on the same raw read dataset. Metagenome assemblies used in Chapter 2 were assembled with SOAPdenovo (Luo et al., 2012) in 2013 and metagenome assemblies used in Chapter 4 with metaSPAdes (Nurk et al., 2017) in 2017. During this time improvements in memory availability and metagenome assembler algorithms have led to an increase in both average contig length (upper panel) and total assembly size (lower panel). In the upper panel all contigs (from largest to shorter contigs) that make up the first 35 Mbp of the assembly are shown. For the selected metaSPAdes assemblies these are ∼300 contigs and for the SOAPdenovo assemblies ∼800 contigs. The lower panel shows the total assembly size of all contigs >500 bp. The black area of the bars highlights the 1000 longest contigs. The graph was visualized using Icarus (Mikheenko et al., 2016).

Even though Chapter 3 provided a good inventory of marine PULs from isolates, these isolates rarely equalled environmentally abundant organisms during our spring phytoplankton sampling time points. Only four of the 53 isolates represented *Bacteroidetes* species detected during 2010 to 2012 bloom events, three of which belonged to the same species. These were the *Polaribacter* isolates Hel1_33_49, Hel1_33_78 and Hel1_33_96 and the *Formosa* isolate Hel1_33_131, which was in more detail functionally described in Appendix 2. This bias in isolation could be based on a reduced physiological flexibility and thus a high specialization on certain substrates of the free-living *Bacteroidetes* (Swan et al., 2013, Woyke et al., 2009), for which isolation requirements could be difficult to determine and achieve. Thus the study of PULs in uncultured bacterioplankton is an important step towards elucidating environmentally relevant PULs. This I was able to achieve in Chapter 4 where we retrieved PULs from the Helgoland spring phytoplankton bloom metagenomes and could subsequently link these to distinct *Bacteroidetes* species. Compared to the PUL types described in Chapter 3 I did not detect entirely new highly abundant PULs in the MAGs. Instead, the PULs detected in the metagenomes in Chapter 4 rather represented a subset of the PUL types described in Chapter 3. Among these were PULs targeting laminarin/β-glucans, α-glucans, mannose-rich substrates, sulphated xylans and alginate that made up the majority of metagenomic PULs (Chapter 4, Fig. 4.3 and Fig. 4.4). Four of these substrate classes were described as expressed during the course of the bloom in Chapter 3 using SusC or SusD expression as an indicator for overall PUL expression. This could be verified by the comparison to the metagenome-derived SusC and SusD sequences in Chapter 4. Additionally the expression of mannose-rich substrates could be confirmed using metagenomic SusC and SusD sequences and the metaproteome data (data not shown). Thus all major polysaccharide substrates for which we detected PULs showed indications of expression based on expressed SusC or SusD proteins in the accompanying metaproteome data. However, it has to be mentioned that the accompanying metaproteome data did not have as high resolution, both regarding sequencing depth and temporal resolution, as our metagenomic data. Thus an interpretation of the metaproteome data on a time perspective was beyond the scope of this study. Nevertheless, the data indicates that all five major substrate classes targeted by metagenomic PULs were of relevance during the spring phytoplankton blooms. Monosaccharide composition of the total combined carbohydrates detected during the 2010 phytoplankton bloom would support these findings as they were dominated by glucose and the co-eluting mannose and xylose (Sperling et al., 2017). This could hint towards the presence of laminarin and α-glucan (both consisting of glucose monomers) as well as mannose-rich substrates and sulphated xylans during the 2010 phytoplankton bloom.

Of the five major substrate classes detected in Chapter 4, α-glucan, alginate and especially laminarin seem to be important substrates for *Bacteroidetes* in the North-Sea. These made up 115/400 PULs in isolates and 84/131 in MAGs. Alginate though was less frequently detected in metagenomic PULs from the free-living *Bacteroidetes* during spring bloom times and was mostly restricted to the NS5 marine group and *Polaribacter* species (Chapter 4, Fig. 4.3). Thus, laminarin and α-glucan are most likely the major relevant substrates, allowing the *Bacteroidetes* community to thrive during phytoplankton blooms. Both laminarin and α-glucan PULs were the most constantly expressed PULs during the studied phytoplankton blooms (Chapter 3, Fig. 3.6). This was also reflected in the presence of PULs targeting these substrates throughout our entire sampling period (Chapter 4, Fig. 4.6). Laminarin and α-glucan PULs were present in abundant *Bacteroidetes* species at pre-bloom, (mid-) bloom and post-bloom time points. They were the most widely distributed PULs among the abundant *Bacteroidetes* species, and present in 14 and 12 out of 27 abundant *Bacteroidetes* species, respectively (Chapter 4, Fig. 4.5). For comparison, PULs targeting alginate were detected in two, and sulphated-xylans and mannose-rich substrates both in five of the abundant *Bacteroidetes* species.

The fact that the PUL repertoire of the uncultured *Bacteroidetes* was restricted to five substrate classes highlights that not only the *Bacteroidetes* biodiversity during algal blooms is constrained, but also their associated PUL repertoires. Since these PUL types were already detected in the *Bacteroidetes* isolates in Chapter 3, it is now possible to investigate the system function-wise. For this, isolate strains with the same or similar PULs compared to the PULs detected in uncultured *Bacteroidetes* could be functionally described including their targeted algae-derived substrates.

5.4 Putative origin of polysaccharide substrates

Laminarin as the major storage polysaccharide in diatoms is one of the most abundant polysaccharides in the ocean, reaching up to 5 to 15 Pg annually produced laminarin (Alderkamp et al., 2007b). Thus it plays an important role during diatom-dominated phytoplankton blooms. It has both been detected in the 10-µm POM and 3-µm POM fraction during a phytoplankton bloom at the North Sea island of Helgoland (Becker et al., 2017). Though in the DOM fraction it regularly does not get detected (Silvia Vidal Melgosa, personal communication), probably due to the high turnover of these rather simple substrates, targeted by a variety of species. Especially *Bacteroidetes* PULs, but also similar gene clusters in *Gammaproteobacteria*, encode for the rapid uptake of larger oligosaccharides or smaller polysaccharides into the periplasm. This has been confirmed for members of the *Bacteroidetes* and *Gammaproteobacteria*

in the Atlantic Ocean, which used a selfish uptake mode for laminarin and other substrates (Reintjes et al., 2018). Uptake could already be detected after 5 minutes of incubation with fluorescently labelled laminarin (Reintjes et al., 2018), which indicates a high turnover rate. Apart from phytoplankton, polysaccharides are also released by the bacterioplankton community. These could be the α-glucan glycogen as the major storage compound of some bacteria (Preiss, 1984) or also bacterial exopolysaccharides (Zhang et al., 2015). Glycogen usage by the *Bacteroidetes* could thus explain the high prevalence of α-glucan PULs. Though these could as well target starch, the storage polysaccharide of red and green algae (Gobet et al., 2018). PULs targeting these two prevalent substrates were also detected in fosmids from two contrasting North Atlantic sites (Appendix 1, Fig. 1.3 and Fig. 1.4). While laminarin PULs were detected at both sites, an α-glucan PUL could only be detected at North Atlantic Subtropical site where algae were less abundant compared to the Boreal Polar site. This highlights the global importance of these two substrates.

Apart from the more prevalent substrates, PULs targeting mannose-rich substrates and sulphated xylans were detected in the *Bacteroidetes* isolates in Chapter 3, metagenomic PULs in Chapter 4 and the North Atlantic fosmids investigated in Appendix 1, which also suggest a global importance of these substrate classes. While laminarin and α-glucan PULs were constantly detected in abundant *Bacteroidetes* MAGs throughout the entire phytoplankton blooms, PULs targeting mannose-rich substrates and sulphated xylans were only detected in species during (mid-) bloom and post-bloom time points. Furthermore their phylogenetic distribution was restricted to the *Aurantivirga*, *Formosa* and *Polaribacter* MAGs, which were the clades that were most diverse in their targeted substrate range. Due to the expression of PULs targeting these substrate classes, they probably play a role especially during and after the phytoplankton blooms. However, for these substrates it remains elusive where they derive from and how relevant they are regarding concentrations during bloom times. The origin of mannose-rich substrates could be diatom cell walls, as described for *Phaeodactylum tricornutum*, where the major cell wall polysaccharide was described as a sulphated glucuronomannan (Le Costaouëc et al., 2017). Likewise other diatom cell wall extracts were dominated by mannose, ranging from 6.5 to 80.1 mol%, and also contained xylose that made up a fraction between 0 and 38.9 mol% of the insoluble organic cell walls in different diatom species (Gügi et al., 2015). Furthermore, diatom and bacterial exopolysaccharides, that are usually anionic and rich in sulphate compounds (Passow, 2002), can contain xylose (Gügi et al., 2015). Xylan hydrolysis has been detected in various regions of the ocean, including the North Atlantic stations described in Appendix 1 (Arnosti et al., 2012). Unfortunately, exact structures of these substrates are unknown, especially those abundant during phytoplankton blooms. This hampers isolation studies. Also PULs, targeting

these substrates, have not yet been experimentally verified. Thus it remains to be elucidated how relevant these two substrates are during spring phytoplankton blooms at Helgoland.

5.5 Different clades with similar substrate ranges - niche specialization or substrate sharing?

Apart from the wide distribution of similar types of PULs among species and the relevance of their targeted substrates, there remains the question of how bacteria that contain similar functions respond to the presence of certain polysaccharide substrates. In Chapter 4 PUL repertoires of the three major clades *Formosa*, *Polaribacter* and *Aurantivirga* did not differ much and all contained similar types of PULs targeting laminarin, α-glucans, mannose-rich substrates and sulphated xylans. Despite the intra-genus variation of their substrate spectra, Mash-clusters 8 (*Formosa*) and 40 (*Polaribacter*) showed distinct abundance patterns but quite similar PUL repertoires, suggesting that the same substrate classes are targeted (Chapter 4, Fig. 4.6). Whether the abundance of these individual species over the course of the bloom is determined by single substrates, a combination of them or whether it is dependent on other factors can until now not be answered with certainty. There are several factors that could determine differences in the species response patterns. These could be, for example, the structural variability in substrates released by the phytoplankton community, combined with specificities of the uptake system and glycan binding proteins for certain substrate types. Furthermore, there could be other environmental triggers influencing the success of one species over another or more stochastic factors, such as the cell numbers at bloom initiation of the distinct species.

Especially with the highly abundant and widespread laminarin/β-glucan targeting PULs, it would be interesting to elucidate more in-depth regulation systems that determine why one or the other species succeeds. Coupled protein and polysaccharide metabolism has been suggested as an advantage for *Formosa* Hel1_33_131 in Appendix 2, where under laminarin incubations both CAZymes and peptidases were expressed. Although my focus in Chapter 4 was mostly on CAZymes, it is likely that other species use similar mechanisms, as we detected peptidases in close vicinity to PULs also in other species. Overall such a coupled mechanism would be of advantage in an environment rich in organic material released by phytoplankton. Other factors determining distinct species' responses to laminarin could be substrate specificities or preferences of the enzymes involved in degradation. For example, Mystkowska et al. (2018) structurally resolved and biochemically analysed a SusD-like protein from *Gramella* MAR_2010_102

and found selective and preferential binding to specific laminarin types. As SusD-like proteins are involved in the uptake of substrate to the periplasm this could be a crucial factor determining which substrates are taken up and eventually which species could have advantages over others based on the substrate variant present. The SusD-like protein from *Gramella* MAR_2010_102 showed preferential binding to laminarin types with a high degree of β-1,6-linkage types or pustulan, a linear β-1,6-glucose polysaccharide (Mystkowska et al., 2018). A SusD-like protein from Mash-cluster 12 (*Cryomorphaceae*) from a laminarin PUL with similar CAZyme composition to the *Gramella* MAR_2010_102 PUL showed high sequence similarity (>40% amino acid identity) to the *Gramella* MAR_2010_102 SusD-like protein and was expressed during the early 2010 bloom, where Mash-cluster 12 reached high abundances (Chapter 4, Fig. 4.10). Assuming that the SusD protein of Mash-cluster 12 has similar substrate specificities as the *Gramella* MAR_2010_102 SusD protein, I hypothesize that during the early 2010 bloom laminarin with a high degree of β-1,6-linkage types might have been a factor influencing the abundance of Mash-cluster 12. Taken together, these specific interactions could be a factor determining the abundance of distinct species. As a matter of fact though, these specific interactions are very laborious to determine as in depth biochemical studies are needed to determine functions and specificities of individual proteins.

Despite these specific interactions functional redundancy for degradation of certain types of polysaccharides has been reported in several studies (D'Ambrosio et al., 2014, Teeling et al., 2012). This means that enzymatic functions for degrading certain polysaccharides are shared between several organisms. How structural preferences play a role in this can often not be addressed, as in-depth knowledge about polysaccharide structure and enzyme specificities is lacking. For example, laminarin degradation is widespread in several organisms. But how individual phytoplankton species' laminarin structure influences fine regulation of preferential uptake is hard to determine without the knowledge of detailed structures of phytoplankton derived polysaccharides. Additonally, Sarmento et al. (2016) showed that the quantity of phytoplankton DOM was more important than the quality. Bacterial communities that were supplied with high concentrations of phytoplankton-derived DOC had a lower specialization index for a certain phytoplankton DOC as compared to when low concentrations were supplied. Under these conditions only a small number of specialists were effectively incorporating the phytoplankton OM (Sarmento et al., 2016). Other factors influencing preferential substrate uptake could be hard wired substrate specificities as detected in gut *Bacteroidetes* (Suncil et al., 2017). But unlike marine free-living bacteria, gut *Bacteroidetes* have a constant supply of similar substrates, for which developing ecological niches or a hard-wired substrate specificity might be a factor in regulation of substrate usage. Whether this is

plausible for marine *Bacteroidetes* that are in a more dynamic environment is questionable.

Taken together, in the broader sense the *Bacteroidetes* species described in this thesis occupy different substrate niches. This becomes especially clear when looking at the narrow PUL repertoires of the *Cryomorphaceae*, the NS5 marine group or *Candidatus* Prosiliicoccus and rather broader PUL repertoires of the *Aurantivirga*, *Formosa* and *Polaribacter* species. However, it remains unclear which mechanisms define these substrate niches that organisms occupy and especially if broader substrate spectra are an advantage for the individual species. In general the niche differentiation pressure on the individual species, at least regarding polysaccharide uptake, might not be that high for the free-living marine *Bacteroidetes* as compared to that for *Bacteroidetes* in other environments. But until we begin to understand the individual species' niches in more detail, especially regarding potentially very narrowly defined polysaccharide substrate niches, this remains speculative.

5.6 Outlook

While I was already able to shed light on the general PUL cosmos and especially the relevant PULs from the free-living marine *Bacteroidetes* during spring phytoplankton blooms, there are still some unaddressed research questions that should be a focus of future studies. One such analysis that could not be addressed in this particular work, but might be of importance in the understanding of which species prevail, is the analysis of individual strains so far hidden on the species-level analysis of MAGs. In Appendix 3 the contribution of *Candidatus* Prosiliicoccus vernus strains was analysed and revealed that the species consisted of 5 confidently predicted strains that displayed consistent abundance patterns during spring phytoplankton blooms. On the other hand, a *Reinekea* species from the *Gammaproteobacteria*, was suggested to be almost clonal during several spring phytoplankton blooms (Avci et al., 2017). While we do not yet know, which of these scenarios is more common among the abundant *Bacteroidetes* species, we already detected variations in the PUL repertoires of closely related strains from the *Polaribacter* genus. The *Polaribacter* isolates Hel1_33_49, Hel1_33_78 and Hel1_33_96, which belong to the same species, carried slightly different PUL repertoires (Chapter 3, Fig. 3.1). In addition, we have indications that such differences in PUL repertoires might also be present in the *Polaribacter* Mash-cluster 35 (data not shown). In general species have a common metabolic machinery that can be supplemented with multiple additional functions present only in a sub-population of strains. Here it would be of special interest if CAZymes or PULs are part of the add-on functions of a certain subpopulation, which would suggest resource partitioning, or whether they are contained in the common metabolic machinery allowing for the coexistence of closely related strains.

Furthermore, it would be of interest to analyse the expression levels of the abundant PULs in the *Bacteroidetes* community during phytoplankton blooms. Although we already know that the five major substrates are expressed, the metaproteome data lacks temporal resolution. Thus a fine-scale analysis using deeply sequenced metatranscriptomes could be of high value in better understanding a temporal, and maybe even successional degradation of phytoplankton derived OM during the course of a bloom. In this respect it would also be of great value to elucidate more in-depth functions for relevant PULs of the community. These are especially PULs targeting mannan-rich substrates and sulphated xylans. While MAGs are fine for the predictions of functions, an actual test of hypothesis can only be performed with isolates that carry similar types of PULs. For these tests though, it is of importance that the polysaccharides used resemble the environmentally abundant substrates, which are often understudied and most likely not commercially available. Thus, extracting and purifying environmentally relevant

substrates should be part of these more in-depth studies and could include structural analyses of these substrates.

Another question would be if the PUL repertoires we detected in the North Sea are comparable to PUL patterns detected in other regions of the ocean. It is likely that in regions also influenced by phytoplankton-derived OM, these PUL patterns are comparable, as was already detected for the taxonomic community composition. But whether *Bacteroidetes* communities in the Arctic or in other marine waters with high influx of terrigenous plant material also carry similar PUL repertoires is something that should be investigated to get a better understanding of the global distribution of PULs. The same applies for the particle-associated *Bacteroidetes* communities. These might already display different functions in PULs, or at least different distributions of PULs among organisms.

5.7 Conclusion

With this body of work I was able to shed light on the major free-living *Bacteroidetes* dominating polysaccharide degradation during algal blooms and provide an overview of their functional repertoires. Despite the complexity and the dynamism of phyto- and bacterioplankton communities during spring blooms, we detected a limited number of *Bacteroidetes* clades that prevail during bloom conditions. These clades were dominated by a few recurrently abundant species that carried a limited number of abundant PULs. Thus also the set of polysaccharides targeted by these organisms is limited and most likely dominated by laminarin and α-glucan-like substrates. From this more holistic view of relevant PULs during phytoplankton blooms that we gained, we can now start to study individual species and their PULs in more detail again. The fact that the free-living *Bacteroidetes* PUL repertoires did not reveal entirely new functional potential, but instead resembled a subset of PULs detected in 53 *Bacteroidetes* isolate genomes, is of great value for these future studies. It means that even though we do not have the representative isolates at hand, we have isolates with PULs that reflect similar functions to the ecologically relevant key-players. Studying these individual PULs in isolates could lead to a more detailed understanding of polysaccharide niches.

Appendix A

Polysaccharide utilisation loci of *Bacteroidetes* from two contrasting open ocean sites in the North Atlantic

Christin M. Bennke[1#‡], Karen Krüger[1‡], Lennart Kappelmann[1‡], Sixing Huang[2], Angélique Gobet[3], Margarete Schüler[4], Valérie Barbe[5], Bernhard M. Fuchs[1], Gurvan Michel[3], Hanno Teeling[1], Rudolf I. Amann[1]

[1]Max-Planck-Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany
[2]Leibniz Institute DSMZ-German Collection of Microorganisms and Cell Cultures, Inhoffstraße 7B, 38124 Braunschweig, Germany
[3]Sorbonne Université, UPMC Univ Paris 06, CNRS, UMR 8227, Integrative Biology of Marine Models, Station Biologique de Roscoff, CS 90074, F-29688, Roscoff cedex, Bretagne, France
[4]University of Bayreuth, Universitätsstraße 30, 95447 Bayreuth, Germany
[5]Laboratoire de Biologie Moléculaire pour l'Étude des Génomes, C.E.A., Institut de Génomique - Genoscope, 2 rue Gaston Crémieux, 91057 Évry cedex, France

[#]present address: Leibniz Institute for Baltic Sea Research Warnemünde, Seestraße 15, 18119 Rostock, Germany

‡These authors made equal contributions

Published in *Environmental Microbiology* (DOI: 10.1111/1462-2920.13429).

Contributions to the manuscript:
Experimental concept and design: 10%
Acquisition of experimental data: 0%
Data analysis and interpretation: 30%
Preparation of figures and tables: 30%
Drafting of the manuscript: 30%

Electronic figure versions and supplementary material are available at https://doi.org/10.1111/1462-2920.13429.

A.1 Abstract

Marine *Bacteroidetes* have pronounced capabilities of degrading high molecular weight organic matter such as proteins and polysaccharides. Previously we reported on 76 *Bacteroidetes*-affiliated fosmids from the North Atlantic Ocean's boreal polar and oligotrophic subtropical provinces. Here, we report on the analysis of further 174 fosmids from the same libraries. The combined, re-assembled dataset (226 contigs; 8.8 Mbp) suggests that planktonic *Bacteroidetes* at the oligotrophic southern station use more peptides and bacterial and animal polysaccharides, whereas *Bacteroidetes* at the polar station (East-Greenland Current) use more algal and plant polysaccharides. The latter agrees with higher abundances of algae and terrigenous organic matter, including plant material, at the polar station. Results were corroborated by in-depth bioinformatic analysis of 14 polysaccharide utilisation loci from both stations, suggesting laminarin-specificity for four and specificity for sulfated xylans for two loci. In addition, one locus from the polar station supported use of non-sulfated xylans and mannans, possibly of plant origin. While peptides likely represent a prime source of carbon for *Bacteroidetes* in open oceans, our data suggest that as yet unstudied clades of these *Bacteroidetes* have a surprisingly broad capacity for polysaccharide degradation. In particular, laminarin-specific PULs seem widespread and thus must be regarded as globally important.

A.2 Introduction

Members of the *Bacteroidetes* phylum are abundant in marine habitats, both in coastal regions (Alonso et al., 2007, Teeling et al., 2012, 2016) and in the open ocean (Gómez-Pereira et al., 2010, Schattenhofer et al., 2009). They occur free-living in the water column as well as attached to particles (Bennke et al., 2013, DeLong et al., 1993). Members of the *Bacteroidetes* are known to be involved in the degradation of high molecular weight dissolved organic matter (HMW-DOM), such as polysaccharides and proteins (Cottrell & Kirchman, 2000, Cottrell et al., 2005, Fernández-Gómez et al., 2013, Thomas et al., 2011). For example, the analysis of the first genome of a marine representative of the bacteroidetal class *Flavobacteriia*, '*Gramella forsetii*' revealed high numbers of peptidase and glycoside hydrolase (GH) genes and thus a high proteolytic and glycolytic potential (Bauer et al., 2006). Similar adaptations have been found in the genomes of other marine *Flavobacteriia*, such as for *Polaribacter dokdonensis* MED152 (González et al., 2008), *Robiginitalea biformata* HTCC2501 (Oh et al., 2009), *Formosa agariphila* KMM 3901 (Mann et al., 2013, Nedashkovskaya et al., 2006) and *Polaribacter* spp. Hel1_33_49 and Hel1_85 (Xing et al., 2015). Metagenomic analyses

also support the view of marine *Bacteroidetes* as specialists for HMW-DOM (e.g. Gómez-Pereira et al. (2010), Teeling et al. (2012, 2016)).

The extent to which *Bacteroidetes* specialize on macromolecular substrates varies considerably. This is reflected in a broad spectrum of CAZyme and peptidase gene frequencies in *Bacteroidetes* genomes. Planktonic *Bacteroidetes* such as '*G. forsetii*' KT0803 tend to have lower (40 GH and 116 peptidase genes; Bauer et al. (2006)) and algae-associated *Bacteroidetes* such as *F. agariphila* KMM 3901 higher CAZyme and peptidase gene numbers (88 GH and 129 peptidase genes; Mann et al. (2013)). *Bacteroidetes* of the human gut are particularly CAZyme-rich with an average of around 130 GHs per genome (El Kaoutari et al., 2013).

The capacity of *Bacteroidetes* for the degradation of polysaccharides is often encoded in distinct polysaccharide utilisation loci (PULs). PULs are operons or regulons of genes that encode the machinery for the concerted detection, hydrolysis and uptake of a dedicated polysaccharide or class of polysaccharides (e.g. Martens et al. (2011)). The starch utilisation system (*susA*-*susG*; *susR*) of the human gut symbiont *Bacteroides thetaiotaomicron* was the first described PUL (Anderson & Salyers, 1989, Shipman et al., 2000). PULs always include an outer membrane transport protein homologous to SusC. This SusC-like protein functions as receptor of the TonB uptake system. In *Bacteroidetes*, this TonB-dependent receptor is usually colocated with an outer membrane lipoprotein homologous to SusD (Bjursell et al., 2006, Cho & Salyers, 2001, Martens et al., 2011, Reeves et al., 1997, Shipman et al., 2000). SusD was shown to bind amylose helices, and to keep starch close to the cell surface of *B. thetaiotaomicron* during degradation (Koropatkin et al., 2008). SusD-like proteins thus define a novel class of carbohydrate-binding proteins and according to current knowledge are unique to the *Bacteroidetes* phylum (Thomas et al., 2011). Within PULs *susC* and *susD* homologs can be colocated with genes coding for GHs, carbohydrate esterases (CEs), carbohydrate binding modules (CBMs), polysaccharide lyases (PLs) and proteins with auxiliary functions. These so-called carbohydrate-active enzymes (CAZymes) are classified in the CAZy database (Cantarel et al., 2009, Lombard et al., 2014). PULs of marine *Bacteroidetes* are frequently found to encode also sulfatases (e.g. Bauer et al. (2006), Gómez-Pereira et al. (2012), Mann et al. (2013), Thomas et al. (2011), Xing et al. (2015)), because in contrast to their land plant counterparts polysaccharides from marine algae are often sulfated (e.g. ulvans, agars, carrageenans, porphyran, fucans). So far most functional studies on PULs have been conducted for land plant polysaccharide-specific PULs in human gut bacteria, for example recently for xyloglucan decomposition by human gut Bacteroidetes (Larsbrink et al., 2014). Only few PULs have been characterized for polysaccharides of marine origin, such as agar/porphyran-specific PULs (Hehemann et al., 2012a) that have been laterally transferred from marine *Bacteroidetes* to *Bacteroidetes* of the human

gut (Hehemann et al., 2010, 2012b) and alginate-specific PULs (Thomas et al., 2012). Notably, alginate induction experiments with *Zobellia galactanivorans* DsiJT demonstrated that its alginate-specific PUL is a genuine operon (Thomas et al., 2012). A recent proteomic study on the coastal marine bacteroidetes '*Gramella forsetii*' KT0803 confirmed expression of proteins encoded by a homologous alginate-specific PUL, and identified an additional laminarin-induced PUL (Kabisch et al., 2014). The latter was also shown to be present and inducible by laminarin in a proteomic study of the coastal marine bacteroidetes *Polaribacter* sp. Hel1_33_49 (Xing et al., 2015). These findings notwithstanding, we still have little knowledge on the PUL repertoire and associated degradation potential of marine *Bacteroidetes*, in particular for those thriving in the mostly oligotrophic open oceans.

In order to gain insights into the genetic capacities for polysaccharide degradation of as yet uncultured open ocean *Bacteroidetes*, we constructed fosmid metagenome libraries from two contrasting provinces of the North Atlantic (Supporting Information Fig. S1 - online) and Table S1 - online) sampled in late September 2006 (Gómez-Pereira et al., 2010). One library of 35,000 fosmids was constructed from surface water taken in the East Greenland Current (station 3) of the Boreal Polar region (BPLR), and a second of 50,000 fosmids was constructed from surface water collected at station 18 (S18) close to the Azores in the North Atlantic Subtropical (NAST) region. Both libraries were screened with a PCR assay targeting the 16S rRNA gene with *Bacteroidetes*-specific primers (CF319 and CF967). A total of 13 (S3) and 15 (S18) fosmids with 16S rRNA genes were identified and sequenced (Gómez-Pereira et al., 2012). Subsequently, we end-sequenced 16,938 S3 and 16,255 S18 fosmids (avg. length: 623 bp). Use of combined results from the end-sequences' tetranucleotide frequency analysis and BLAST and HMMer searches of encoded genes allowed identification of additional fosmids of possible bacteroidetal origin. In a previous study, we reported on the analysis of the first 76 fully sequenced *Bacteroidetes* fosmids (Gómez-Pereira et al., 2012). This analysis revealed that *Bacteroidetes* from both regions had an unexpectedly high capacity for polymer degradation in view of the overall nutrient depletion of open oceans. The analysis also suggested that *Bacteroidetes* in the more oligotrophic southern region might be more adapted towards the degradation of proteins and peptidoglycan than to polysaccharides of algal origin (Gómez-Pereira et al., 2012).

Here, we present the analysis of 174 new bona fide *Bacteroidetes* fosmids from the BPLR (S3: 95) and NAST (S18: 79) of the Northern Atlantic, which extends the initial dataset to a total of 250 fosmids. Re-assembly yielded 226 contigs, which we analysed in terms of peptidases, CAZymes and putative PULs with a special focus on possible polysaccharide substrates, and on the question as to whether differing oceanic provinces select for *Bacteroidetes* clades with different CAZyme repertoires.

A.3 Results and discussion

A.3.1 Characterisation of the dataset

The initial 76 fosmids (S3: 40; S18: 36) were pooled with 154 newly sequenced putative *Bacteroidetes* fosmids (S3: 84; S18: 70) and re-assembled. This way, 107 contigs were obtained from S3 and 95 contigs from S18. Some of the new fosmids were selected for sequencing due to high similarity of their end-sequences to previously sequenced fosmids. Thereby it was possible to extend some of the initial fosmids and to obtain assemblies of up to 85.4 kbp (S3) and 72.2 kbp (S18). In addition, 14 putative PULs were retrieved that were not obtained before. Additional fosmids (S3: 11; S18: 9) were selected in order to possibly extend these PUL-carrying fosmids, which after assembly yielded additional 24 contigs (S3: 14; S18: 10). In summary, the total dataset comprised 250 fosmids (S3: 135; S18: 115) that were re-assembled to 226 contigs (S3: 121; S18: 105) of 8.8 Mbp (S3: 4.7 Mbp; S18: 4.1 Mbp).

Based on gene content, 96% (S3) and 92% (S18) of the contigs affiliated with *Bacteroidetes*. Thus, the error of selecting *Bacteroidetes* fosmids based on information from combined ∼1.4 kbp Sanger sequenced end-sequences was comparable to the ∼5% error of PCR-based screening (Gómez-Pereira et al., 2012). Sequenced non-*Bacteroidetes* contigs affiliated with the *Planctomycetes–Verrucomicrobia–Chlamydia*-cluster and with *Proteobacteria*. On class level 95% of the *Bacteroidetes* contigs affiliated with *Flavobacteriia* at S3 and 94% at S18, of which 96% affiliated with the family *Flavobacteriaceae* at both stations. On genus level (Fig. 1.1) S3 contigs affiliated most frequently with *Polaribacter* (39)%, *Flavobacterium* (13%), *Dokdonia* (6%) and *Gramella* (6%), whereas S18 contigs most frequently affiliated with *Dokdonia* (25%), *Leeuwenhoekiella* (17%), *Flavobacterium* (11%) *Robiginitalea* (8%), *Polaribacter* (7%) and *Croceibacter* (7%). The numbers obtained for *Polaribacter* are in good agreement with *in situ Polaribacter* abundances that were previously determined by catalysed reporter deposition fluorescence in situ hybridisation (CARD-FISH) of the same samples (Gómez-Pereira et al., 2010). However, *Gramella* was not detected using CARD-FISH at both stations, and *Leeuwenhoekiella* and *Dokdonia* were only detected at station 18 with abundances below 1% (Gómez-Pereira et al., 2010). Such discrepancies are expected, as our taxonomic affiliations are based on gene BLASTp and HMMer database similarity searches and thus biased towards publically available sequenced *Flavobacteriia*.

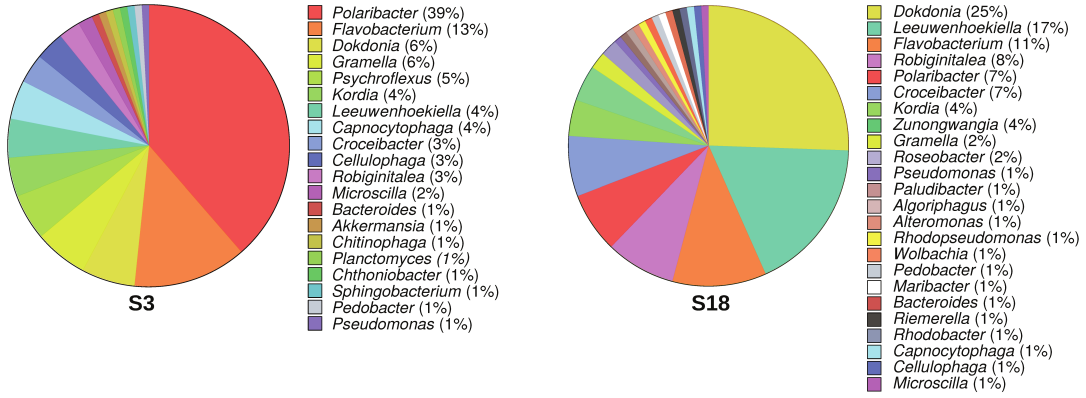


Figure 1.1: Genus-level taxonomic affiliation of contigs obtained from reassembled fosmid sequences of metagenomic libraries from station S3 (121 contigs) and S18 (105 contigs).

A.3.2 *Bacteroidetes*' peptidases and CAZymes

The numbers of predicted peptidase genes agree by and large with previously published values of Gómez-Pereira et al. (2012). The dataset from NAST station 18 had significantly higher peptidase frequencies than the one from BPLR station 3 (Fig. 1.2 A), and in both cases, peptidase frequencies exceeded those of glycolytic CAZymes (GHs, CEs, PLs) by a factor of 1.7 and 2.9 respectively (S3: 25.5 Mbp^{-1} vs. 14.7 Mbp^{-1}; S18: 37.4 Mbp^{-1} vs. 13.1 Mbp^{-1}; Table 1.1; Fig. 1.2 B and C). Quantitative comparison of peptidase and CAZyme gene numbers based on automatic predictions may contain a certain amount of error due to the involvement of different databases of different sizes and different E-value thresholds. However, our findings are in agreement with the observation that most sequenced genomes of marine *Bacteroidetes* feature higher numbers of peptidase than CAZyme genes (Fernández-Gómez et al., 2013, Xing et al., 2015). Especially planktonic *Bacteroidetes* with small to average-sized genomes tend towards higher peptidase:CAZyme ratios. In contrast, alga-associated *Bacteroidetes* species feature mostly CAZyme-rich large genomes where the combined number of CAZyme genes can exceed those of peptidase genes (Xing et al., 2015). Thus, higher peptidase to CAZyme ratios would be expected for planktonic *Bacteroidetes* from open ocean sites, in particular at a more oligotrophic site like station 18. Our results indicate that proteins (and amino acids) contribute substantially to carbon and nitrogen uptake in open ocean *Bacteroidetes*. The number of peptidase families was also higher at station 18 (S3: 44; S18: 55) as were the frequencies of some individual families such as C40, C44, M38, M42, S09X, S12 and S54, whereas others such as M16B, M50B, S08A and S45 family peptidases were more frequent at S3 (Fig. 1.2 B).

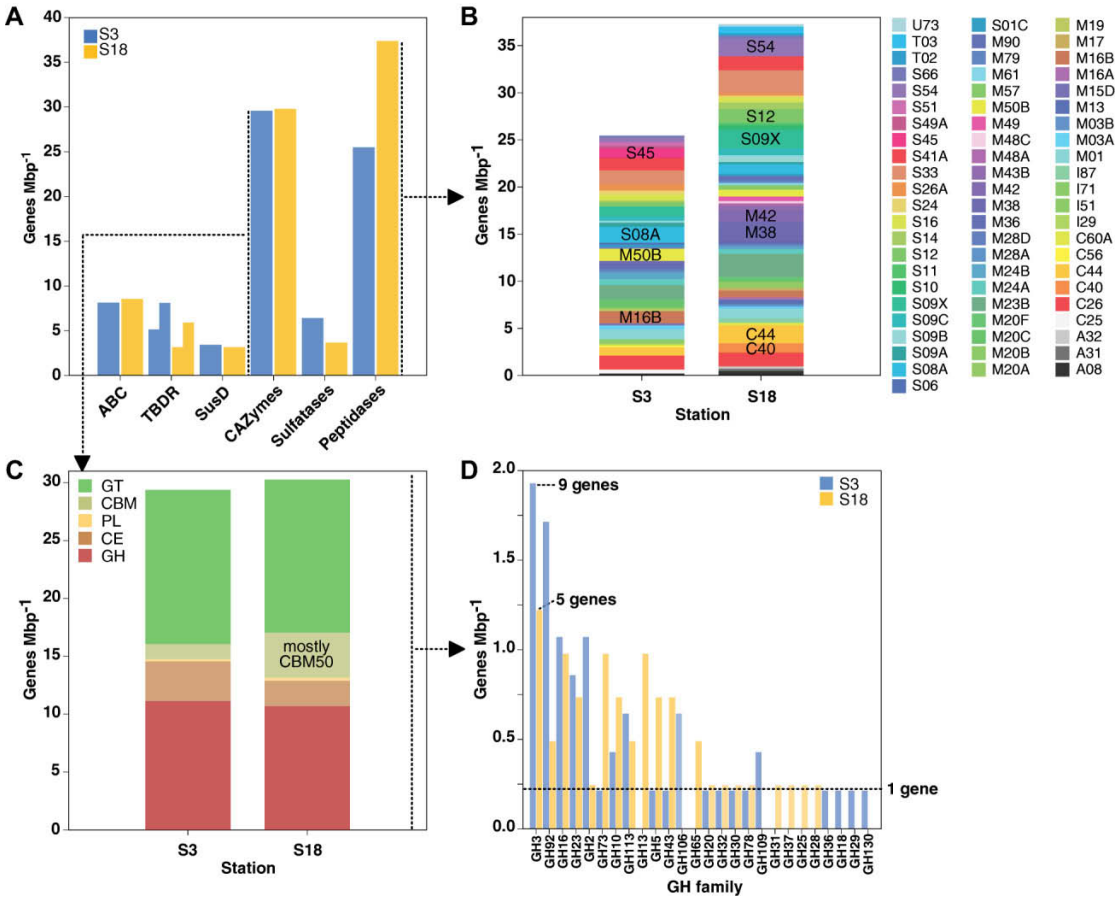


Figure 1.2: Comparison of selected genes on the 121 BPLR station 3 contigs (4.7 Mbp) and the 105 NAST station 18 contigs (4.1 Mbp). A.) Genes for ABC-transporters, TonB-dependent receptors (TBDRs), SusD-like proteins, CAZymes, sulfatases and peptidases. B.) Peptidase families. C. CAZyme classes: glycoside hydrolases (GH), carbohydrate esterases (CE), polysaccharide lyases (PL), carbohydrate-binding modules (CBM) and glycosyltransferases (GT). D.) GH families ordered by decreasing averages.

Genes for polysaccharide binding (*susD*) and GH genes exhibited about equal frequencies in both datasets, while TonB-dependent receptor and sulfatase genes were 1.6- to 1.7-fold more frequent in S3 than in S18 contigs (5.1 vs. 3.2 and 6.4 vs. 3.7 Mbp^{-1}; Table 1.1; Fig. 1.2 A). The latter indicates a higher prevalence of sulfated algal polysaccharides at the polar station S3, which agrees with higher chlorophyll *a* measurements at this station (S3: 0.7 $\mu\text{g l}^{-1}$; S19: < 0.1 $\mu\text{g l}^{-1}$; Gómez-Pereira et al. (2010)). CBM-containing genes were three times more frequent at NAST station 18 than at BPLR station 3 (3.9 vs. 1.3 Mbp^{-1}). Most of them belonged to the peptidoglycan- or chitin-binding CBM50 family (Fig. 1.2 C). This agrees with the analysis of the initial dataset by Gómez-Pereira et al. (2012), who reported higher numbers of peptidoglycan degradation genes at station 18 versus station 3.

Table 1.1: Comparison of frequencies of genes involved in organic matter degradation and uptake between the S3 and S18 contig datasets. [a] HMMer 3 searches against the Pfam v. 28 database with $E \leq 10^{-5}$. Values for ABC transporter and *susD* genes were determined by combining all genes with hits to any of the following profiles: ABC_tran, ABC_membrane, ABC_membrane_2, ABC_membrane_3, ABC_tran_2, ABC2_membrane, ABC2_membrane_2, ABC2_membrane_3, ABC2_membrane_4, ABC2_membrane_5, ABC2_membrane_6, as well as SusD, SusD-like, SusD-like_2 and SusD-like_3. Values for TonB-dependent receptor genes were determined using the TonB_dep_rec profiles from October 2014 and 2012 (values in brackets). Both TonB_dep_rec profiles predicted lower numbers as was suggested by BLAST-based database similarity searches. [b] Combined results of BLASTp searches against the CAZy database, HMMer searches against the Pfam v. 28 database and the dbCAN database with manually adjusted E-value thresholds (Teeling et al., 2016). [c] Batch BLASTp searches against the MEROPS 9.13 database with $E \leq 10^{-4}$ (the database is small, hence $E \leq 10^{-4}$ is considered significant).

Genes/Mbp	S3	S18
ABC transporters genes[a]	8.1	8.5
TonB-dependent receptor genes[a]	5.1 (8.1)	3.2 (5.9)
susD genes[a]	3.4	3.2
CAZymes[b]	29.5	29.7
GHs	11.1	10.7
CBMs	1.3	3.9
CEs	3.4	2.2
PLs	0.2	0.2
GTs	13.5	12.7
GHs + CEs + CLs	14.7	13.1
sulfatase genes[a]	6.4	3.7
peptidase genes[c]	25.5	37.4

Among the most frequent GH families (Fig. 1.2 D), some were about equally represented in contigs from both stations, e.g. GH16 and GH23, whereas for example GH3 (a family comprising diverse functions) and GH92 (a family comprising mostly alpha-mannosidases) were more frequent in the S3 dataset than on S18 contigs (GH3: 1.9 vs. 1.2 Mbp^{-1}; GH92: 1.7 vs. 0.5 Mbp^{-1}). In the less abundant families, GH2 members were found at higher frequencies in S3 contigs (1.1 vs. 0.2 Mbp^{-1}), whereas families GH73, GH5 and GH43 were found at higher frequencies in S18 contigs (GH73: 0.2 vs. 1.0 Mbp^{-1}; GH5: 0.2 vs. 0.7 Mbp^{-1}; GH43: 0.2 vs. 0.7 Mbp^{-1}). Among the GH families with at least two members in either of the datasets, GH106 (0.6 Mbp^{-1}) and GH109 (0.4 Mbp^{-1}) were found only at station 3, GH13 (1.0 Mbp^{-1}) and GH65 (0.5 Mbp^{-1}) only at station 18 and 10 GH families were found at both stations (GH2, 3, 5, 10, 16, 23, 43, 73, 92, 113). Within the entire dataset, 14 putative PULs were detected (Supporting Information S3: 6 and S18: 8; Figs. 1.3 and 1.4) comprising 40 GHs of 17 families. In order to gain insights on possible polysaccharide substrates, we conducted in-depth manual annotation of these PULs (Supporting Information Table S2 - online).

A.3.3 Commonalities between PULs from both stations

GH16 was among the most frequent GH families in the dataset (Fig. 1.2 D). Five GH16 genes were found in four of the putative PULs, namely on contigs VISS3_015 and VISS3_033 from station S3 (Fig. 1.3) and VISS18_021 and VISS18_090 from station S18 (Fig. 1.4). An extracellular location was predicted for all corresponding GH16 proteins (Supporting Information Table S3 - online). The family GH16 comprises various enzymatic activities (Eklöf et al., 2013, Lombard et al., 2014, Michel et al., 2001) and thus phylogenetic analyses are required to determine specificities of GH16 members. Such analyses suggested that the encoded GH16 enzymes are beta-1,3-glucanases usually referred to as laminarinases (Supporting Information Fig. S2 - online). These laminarinases encompass enzymes that act on different types of biologically unrelated beta-1,3-glucans, and only few of these enzymes are specific for genuine laminarin (Labourel et al., 2014). Laminarin (a beta-1,3-glucan with occasional beta-1,6 branching) is the storage polysaccharide of brown algae (Michel et al., 2010) and of diatoms (known as chrysolaminarin; Beattie et al. (1961)) and belongs to the most abundant polysaccharides on Earth. Four of the GH16 enzymes (S3: ORFs VISS3_015_23, VISS3_033_04; S18: ORFs VISS18_021_09, VISS18_090_12) belong to a clade that contains $ZgLamA_{GH16}$ from Z. galactanivorans DsiJT (Supporting Information Fig. S2 - online). $ZgLamA_{GH16}$ features an extra loop leading to a bent active site that provides high specificity for genuine algal beta-1,3-glucans (Labourel et al., 2014). The fifth GH16 on contig VISS3_015 (ORF VISS3_015_21) clustered with two functionally uncharacterised GH16 from *Flavobacterium* species. It might be distantly related to the clades containing ZgLamB and ZgLamC (Supporting Information Fig. S2 - online), which do not possess the characteristic $ZgLamA_{GH16}$ loop and act on beta-1,3-glucans and mixed-linkage (beta-1,3-1,4) glucans (Labourel et al., 2015). ORF VISS3_015_21 might encode a similar broad specificity beta-glucanase.

The predicted GH16 laminarinase genes in all four PULs are each colocated with a GH3 family gene. Two (ORFs VISS3_015_22, VISS3_033_03) of these four GH3 genes were predicted to code for beta-glucosidases and two (ORFs VISS18_021_08, VISS18_090_13) for beta-glycosidases (Supporting Information Fig. S3 - online). These GH3 enzymes likely hydrolyse terminal beta-d-glucosyl residues from oligo-laminarin.

CAZyme analysis also suggested xylose-rich polysaccharides as potential substrates at both stations. Particularly interesting in this context is the PUL on the extended S3 contig VISS3_016 (VISS3_016 + VISS3_057; Fig. 1.3). This PUL harbours a gene that codes for a putative modular enzyme with an N-terminal sulfatase and a C-terminal GH10 family xylanase (ORF VISS3_016_05). The physical link between these two enzymatic activities indicates degradation of a sulfated

xylose-rich polysaccharide. This is further supported by presence of two sulfatase genes (ORFs VISS3_016_02, VISS3_016_03) and a predicted GH3 family beta-1,4-xylosidase gene (ORF VISS3_016_04; Supporting Information Fig. S3 - online) in this PUL. Xylan metabolism GHs and sulfatases were also predicted in the PUL on contig VISS18_012 (Fig. 1.4): a GH3 family xylan 1,4-beta-xylosidase (ORF VISS18_012_31; Supporting Information Fig. S3 - online), a GH10 family endo-beta-1,4-xylanase (ORF VISS18_012_17) and a GH43 family beta-xylosidase/alpha-l-arabinofuranosidase (ORF VISS18_012_15). This PUL is more complex than the PUL from VISS3_016, as it codes for additional CAZymes: a GH10 (ORF VISS18_012_32), a GH30 (ORF VISS18_012_22), a PL9 polysaccharide lyase (ORF VISS18_012_18), a CE1 carbohydrate esterase (ORF VISS18_012_09) and no less than six sulfatases (ORFs VISS18_012, _19, _20, _21, _29, _30, 32). Adjacent to the GH43 and GH10 genes a xyloside transporter gene (*xynT*) was predicted, which might transport xyloside into the cytoplasm, where it is further converted to xylulose-5-phosphate via xylose and xylulose. The respective genes *xylA* (xylose-isomerase) and *xylB* (xylulose-kinase) were identified downstream of the PUL. Neutral xylan occurs in red and green algae (Popper et al., 2011), and sulfated xylan has been found in the red macroalga *Palmaria palmata* (Deniaud et al., 2003). Likewise, exopolysaccharides (EPS) of some diatoms and bacteria are sulfated and rich in xylose.

The predicted capacity to decompose laminarin and xylan or xylose-rich polysaccharides agrees with findings of Arnosti et al. (2012), who demonstrated the potential for laminarin and xylan hydrolysis using fluorescein-labelled polysaccharides at BPLR station 3 and NAST station 19 of the same cruise (Supporting Information Fig. S1 - online). Laminarin and sulfated xylose-rich polysaccharides are probably more prevalent at the northern BPLR station 3 than at NAST station 18, as S3 is located at the lower border of the East Greenland Current, which transports cold, low saline, but nutrient- and phytoplankton-rich waters from the Arctic Ocean alongside the eastern coast of Greenland southwards (Bersch, 1995) while the NAST is a typical oligotrophic 'blue' ocean (Longhurst, 2006).

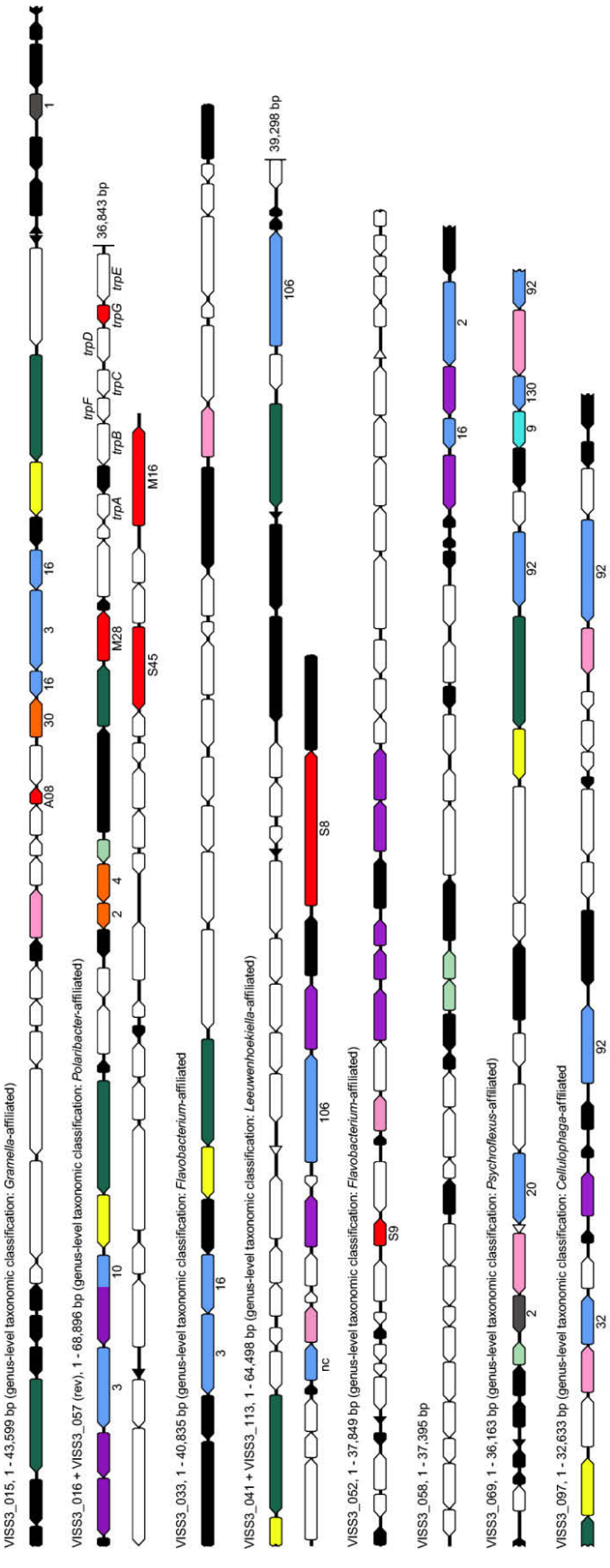


Figure 1.3: PUL-containing contigs from BPLR station 3. Names, presumed taxonomic affiliations, lengths and the gene contents are provided for each contig.

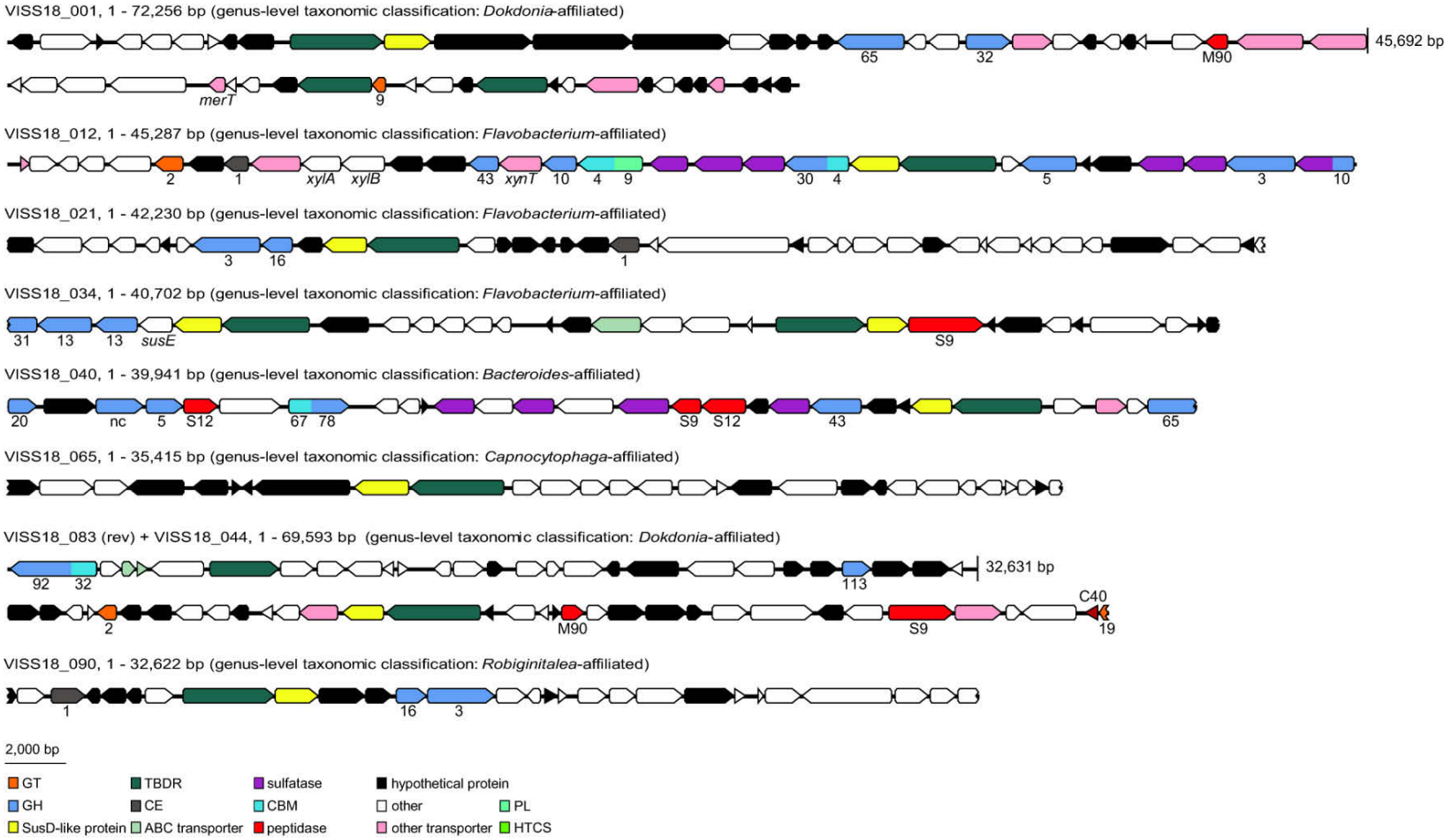


Figure 1.4: PUL-containing contigs from NAST station 18. Names, presumed taxonomic affiliations, lengths and the gene contents are provided for each contig.

A.3.4 Additional PULs from BPLR station 3

The extended contig VISS3_113 (VISS3_041 + VISS3_113; 64.5 kbp) encodes a PUL with two predicted GH106 alpha-rhamnosidase genes (ORFs VISS3_113_01, VISS3_113_16). Alpha-l-rhamnosidases are known to cleave terminal alpha-l-rhamnose from cell wall polysaccharides of plants (in rhamnogalacturonans) and green algae (ulvans, a family of sulfated xylorhamnoglucuronans) (Lahaye & Robic, 2007, Martin et al., 2014, Naumoff & Dedysh, 2012, Popper et al., 2011). This PUL also contains two carbohydrate sulfatase genes (ORFs VISS3_113_14, VISS3_113_17) and a non-classified GH (ORF VISS3_113_10), which could act in synergy with the GH106 enzymes in the degradation of sulfated rhamnose-containing polysaccharides.

Contig VISS3_069 has been classified as *Psychroflexus*-related. Known *Psychroflexus* spp. are psychrophilic, colonize surfaces of sea-ice diatoms (Bowman et al., 1998, Sullivan & Palmisano, 1984) and use a rather narrow range of substrates, possibly obtained from diatom EPS (Malinsky-Rushansky & Legrand, 1996). Contig VISS3_069 harbours a putative PUL with genes involved in xylan metabolism. There is a putative CE2 family acetyl-xylan esterase (ORF VISS3_069_11), involved in deacetylation of xylans and xylo-oligosaccharides, and a protein of unknown function containing a CBM9, a module which has so far been only found in xylanases (Lombard et al., 2014). This PUL also codes for enzymes that likely take part in degradation of mannan or mannose-rich glycoproteins: two putative GH92 family alpha-1,2-mannosidases (ORFs VISS3_069_22, VISS3_069_28), a putative GH20 hexosaminidase (ORF VISS3_069_14), which may act on n-acetylmannosamine (Senoura et al., 2011), and a GH130 family enzyme (ORF VISS3_069_26). The GH130 family comprises beta-1,4-mannooligosaccharide phosphorylase, an enzyme that is involved in a novel mannan catabolic pathway (Senoura et al., 2011). As this PUL lacks any obvious endo-polysaccharidase, it is probably incomplete. Xylans and mannans are part of plant cell walls, but they are also found in red algae (Popper et al., 2011) and they play an important structural role in diatom cell walls (Hecky et al., 1973). However, land plant biomass is a more likely source for non-sulfated xylans and mannans at the BPLR station 3, as the Arctic Ocean water that is transported with the East Greenland Current towards station S3 is sevenfold to 16-fold richer in terrigenous dissolved organic matter than the Atlantic and Pacific Oceans (Benner et al., 2005) and includes high amounts of land plant material such as driftwood (Hellmann et al., 2013).

Contig VISS3_097 has been classified as *Cellulophaga*-related. *Cellulophaga* species are known to be associated with diatoms and macrophytes from cold marine waters, e.g. *C. algicola* (Bowman, 2000). This PUL on this contig contains a sulfatase (ORF VISS3_097_09), two putative GH92 family alpha-1,2-mannosidases

(ORFs VISS3_097_13, VISS3_097_22) and a putative GH32 family levanase (ORF VISS3_097_05). The latter might hydrolyze 2,6-beta-d-fructofuranosidic linkages in 2,6-beta-d-fructans. Fructans (or levans for bacteria) can be synthesized by green algae or bacteria. Bacterial fructans are produced extracellularly and generally composed of beta-2,6-linked fructosyl residues linked to a terminal glucose, such as for example in Lactobacillus and Streptococcus species (Corrigan & Robyt, 1979, Hendry, 1993, Van Geel-Schutten et al., 1999). Therefore, this PUL might be dedicated to the degradation of sulfated EPS containing mannose and/or fructose residues.

Contig VISS3_052 did not contain the characteristic *susCD* gene pair of PULs, but it might constitute a partial PUL as indicated by the presence of five sulfatases (ORFs VISS3_052, _20, _21, _22, _25, _26), one putative sugar transporter, a carbohydrate kinase, two sugar isomerases and a fructose-1,6-bisphosphate aldolase gene. Two genes (ORFs VISS3_052_12, VIS3_052_24) on this contig share remote similarities with GHs, but no sufficient similarity to be annotated as such.

A.3.5 Additional PULs from NAST station 18

The PUL on contig VISS18_034 encodes two predicted GH13 family alpha-amylases (ORFs VISS18_034_02, VISS18_034_03) from distinct subfamilies (Stam et al., 2006). In phylogenetic reconstruction (Supporting Information Fig. S5 - online), ORF VISS18_034_03 clustered with two GH13_7 subfamily alpha-amylases from Thermococcus species, suggesting lateral gene transfer (LGT). In contrast, ORF VISS18_034_02 belongs to the GH13_20 subfamily that includes cyclomaltodextrinase, neopullulanase and maltogenic amylase. In comparison to classical alpha-amylases, these enzymes, and also ORF VISS18_034_02, feature an extra domain that participates in dimer formation (Lee et al., 2002, Stam et al., 2006). Cyclomaltodextrinases effectively hydrolyze cyclomaltodextrin, a circular sugar derived from starch degradation, whereas the degradation of starch and pullulan is less effective (Lee et al., 2002). The PUL also contains a putative GH31 family alpha-glucosidase (ORF VISS18_034_01), which hydrolyzes the oligosaccharides released by alpha-amylases. Therefore, this PUL likely targets an alpha-1,4-glucan, which is for example produced by some bacteria as storage compound during stationary growth (Field et al., 1998, Preiss, 1984). Usage of bacterial polysaccharides by Bacteroidetes may be of higher relative importance at NAST station S18 as algae were less abundant than at the BPLR station S3 (Supporting Information Table S1 - online).

The PUL on contig VISS18_001 encodes a predicted GH32 family beta-fructofuranosidase (ORF VISS18_001_23). These enzymes hydrolyze terminal non-reducing beta-d-fructofuranoside residues in beta-d-fructofuranosides (for instance sucrose). The PUL also encodes a putative GH65 family maltose phosphorylase (ORF VISS18_001_20). These enzymes add phosphate to maltose, resulting in d-glucose and beta-d-glucose-1-phosphate.

The PUL on contig VISS18_040 contains enzymes involved in starch and sucrose metabolism, such as a predicted GH65 family maltose phosphorylase (ORF VISS18_040_30) and GH43 family xylosidase/arabinofuranosidase (ORF VISS18_040_21). This PUL also contains a putative GH20 beta-n-acetylhexosaminidase (ORF VISS18_040_01), a GH5_13 glycoside hydrolase that may target beta-mannan (ORF VISS18_040_04; Supporting Information Fig. S4 - online; Aspeborg et al. (2012)), a putative GH109 alpha-N-acetylgalactosaminidase (ORF VISS18_040_13) and a GH of unknown specificity (ORF VISS18_040_03). Furthermore, the PUL contains a predicted GH78 family alpha-l-rhamnosidase (ORF VISS18_040_07) and four sulfatases (ORFs VISS18_040_12, _14, _16, _20; family S1 and subfamilies 8, 9, 16, 20; Supporting Information Table S2 - online), similar to the PUL of S3 contigs VISS3_113 + VISS3_041. These enzymes may be involved in the hydrolysis of sulfated polysaccharides (e.g. algal ulvans).

Annotations did not provide sufficient information for hypotheses on possible substrates for PUL-containing contigs VISS18_065 and VISS18_083 + VISS18_044 (Fig. 1.4). Contig VISS18_083 + VISS18_044 harbours a putative CBM32-containing GH92 alpha-1,2-mannosidase gene (ORF VISS18_083_32) and a putative CE10 gene, whereas contig VISS18_065 featured a *susC-susD* gene pair, but otherwise no matches to the CAZy database.

A.3.6 Comparative analysis of PULs

Some of the analysed PUL-containing contigs share regions of high DNA sequence similarity. Such homology was observed not only among PULs from the same sampling station, but also between PULs from both stations. For instance, two regions of contigs VISS3_052 and VISS3_041 + VISS3_113 are highly similar (Supporting Information Fig. S6 A - online). The first (VISS3_052: 12.89–15.82 kbp; VISS3_041 + VISS3_113: 51.28–54.32 kbp) comprises two (l-arabinose isomerase, sulfatase) and the second (VISS3_052: 23.54–33.45 kbp; VISS3_041 + VISS3_113: 12.00–22.81 kbp with

a small insertion at 19.54–20.10 kbp) seven genes (membrane protein, l-lactate dehydrogenase, carbohydrate kinase, sugar isomerase, short-chain dehydrogenase, fructose-1,6-bisphosphatase, transcriptional regulator - GntR family). Likewise, contigs VISS3_069 (26.60–28.87 kbp) and VISS3_097 (13.11–15.36 kbp) share a region of high DNA similarity that encodes a GH92 family enzyme (Supporting Information Fig. S6 B - online).

BPLR contigs VISS3_016 + VISS3_057 and the NAST contig VISS18_012 have a highly similar region harbouring homologous GH3 family genes, neighboured by three non-homologous sulfatase genes on both contigs (Supporting Information Fig. S6 C - online). Similarly four of the putative laminarinase-containing contigs from both stations (BPLR: VISS3_015, VISS3_033; NAST: VISS18_021, VISS18_090) exhibited a high level of sequence conservation (Supporting Information Fig. S6 D - online).

Comparative genomics suggests that LGT is frequent among human gut Bacteroidetes (Coyne et al., 2014), including exchange of CAZymes and entire PULs (Hehemann et al., 2010, 2012b, Martens et al., 2014). We found homologous regions in more than half of the PULs that we analysed, which suggests that such LGT events might occur also rather frequently among marine Bacteroidetes. These analyses also suggest that parts of PULs can be laterally transferred and act as recombination modules in PUL evolution. Whether entire PULs can be laterally transferred from *Bacteroidetes* to non-*Bacteroidetes* is an open question. The fact that susD genes have so far only been found in *Bacteroidetes* suggests some type of recombination barrier that prevents establishment of a *Bacteroidetes*-like SusC–SusD interaction in other phyla.

A.4 Conclusion

Although a fosmid-based approach is more laborious and comes at a much higher cost per base than a shotgun metagenome approach, it has the advantage of being targeted, as fosmid libraries can be end-sequenced and screened for clones from dedicated taxa, and it is guaranteed to provide sequences that are long enough for the study of larger gene arrangements such as PULs.

In their initial study, Gómez-Pereira et al. (2012) concluded that *Bacteroidetes* at the BPLR station S3 were richer in polysaccharide degradation genes than at NAST station S18, who in turn had higher peptidase gene frequencies. Our analysis of the extended fosmid dataset confirmed higher peptidase frequencies at NAST, but higher CAZyme frequencies at the BPLR station could not be substantiated. In the present study, we particularly focused on in-depth manual annotation of CAZymes and CAZyme genome

clusters so as to provide a more substantial idea of their activity beyond automated assignment to diverse enzyme families based on bioinformatic tools. Although frequencies of CAZymes were not different between stations, composition of the respective CAZyme sets clearly was, and the prevalence of sulfatases was notably higher at BPLR than at NAST. This agrees with results of Arnosti et al. (2012), who found the microbial community at the BPLR station 3 to be capable of degrading the sulfated polysaccharides chondroitin and fucoidan at a faster rate than at the NAST station 19 of the same cruise. Fittingly, colocalisations of sulfatases and GHs were found on six of the 121 S3 contigs (including fosmids S3-860 and S3_DL_C5 reported by Gómez-Pereira et al. (2012) not shown in Fig. 1.3), but only on two of the 105 S18 contigs. Sulfated polysaccharides are produced in large quantities by marine algae, which were more abundant at BPLR station 3 than at NAST station 18 (Supporting Information Table S1 - online). This means that the relative contribution of carbon from amino acids/peptides and from non-algal organic matter was higher at NAST station 18 than at BPLR station 3, which is reflected in (i) higher peptidase frequencies, (ii) higher frequencies of GHs for the hydrolysis of bacterial or animal alpha-1,4-glucans and (iii) a higher prevalence of CBM50 genes that might cleave either bacterial peptidoglycan or animal chitin. Conversely, higher frequencies of GH92 mannosidases at BPLR station 3 support a higher importance of algal- and (see below) plant-derived polysaccharides at this station.

PUL comparisons indicated that sulfated xylan-rich polysaccharides and algal laminarin are possibly among the more frequent polysaccharide substrates for open ocean *Bacteroidetes*. One PUL at each station contained putative xylan-specific genes and sulfatases, suggesting xylan-rich polysaccharide of marine origin as substrates (e.g. from algal EPS). Four out of the 14 PULs in our dataset were likely laminarin-specific. Similar PULs have been identified in other members of the *Bacteroidetes* (Kabisch et al., 2014), suggesting that such PULs are widespread and of global importance. This underpins the importance of laminarin as substrate in the marine realm, also in the open ocean.

The dataset presented in this study demonstrates that the CAZyme repertoire of *Bacteroidetes* in open ocean sites such as the BPLR station 3 and the NAST station 18 is diverse and even comprises families such as GH10, 43, 78 and 106 that have been suggested to be characteristic for *Bacteroidetes* feeding on land plant biomass (Kolton et al., 2013). Marine plants and algae produce some polysaccharides that are usually found in terrestrial plants. Rhamnogalacturonans and xylans for example are present in land plant hemicelluloses, but rhamnogalacturonans are also constituents of pectins in marine angiosperms, and xylans have also been found as a form of cell covering in some marine algae (Okuda, 2002) or as cell wall components in some diatoms (Murray et al., 2007, Wustman et al., 1998). Presence of the above mentioned GH families, at least at BPLR station 3, may, however, be most likely explained by the station's location within

the East Greenland Current, that transports ample terrigenous organic matter including plant material (Benner et al., 2005, Hellmann et al., 2013). This would also explain the finding of a partial PUL with xylan- and mannan-specific genes without sulfatases at this station.

At oligotrophic open-ocean sites, algal and bacterial polysaccharides are produced in much lower amounts than at eutrophic sites. Therefore, these energy-rich compounds are particularly valuable at open ocean sites, and consequently heterotrophic bacteria exist at such sites that can consume these polysaccharides when they become available. As in other habitats, the *Bacteroidetes* seem to play a key role in such turnover of complex organic matter also in open oceans. It will be up to future systematic studies to inventory the PUL repertoire of marine *Bacteroidetes* in a comprehensive manner and to explore, which individual PULs are ubiquitously distributed and thus most important in marine habitats.

A.5 Experimental procedures

A.5.1 Study sites and fosmid library preparation

Samples were taken in the North Atlantic Ocean during the VISION cruise MSM03/01 on board of R/V Maria S. Merian in September 2006 (Supporting Information Fig. S1 - online and Table S1 - online). Fosmid metagenome libraries were constructed from surface water of two contrasting oceanic provinces. Samples from station 3 (S3) were collected in the Boreal Polar province (65°52.64′ N, 29°56.54′ W) and samples from station 18 (S18) in the NAST province (34°04.43′ N, 30°00.09′ W). Libraries of 35,000 (S3) and 50,000 (S18) fosmids were constructed for both sites. Subsequently, 16,938 (S3) and 16,266 (S18) high-quality end-sequences were generated by sequencing inserts from both sites using the Sanger technique. Details have been described elsewhere (Gómez-Pereira et al., 2010, 2012).

A.5.2 Selection and sequencing of fosmids

End-sequences were mapped on the 76 previously sequenced fosmids from both libraries in order to detect connecting fosmids. Using a sequence identity threshold of 94.5% or higher, 43 (S3) and 27 (S18) connecting fosmid candidates were identified. Twenty of these had the potential to prolong partial PULs on the previously sequenced fosmids.

These were sequenced at LGC Genomics (LGC Genomics GmbH, Berlin, Germany) using the 454 FLX Ti platform and assembled using Newbler. Further 104 putative *Bacteroidetes* fosmids were selected based on BLASTx hits of end-sequences to the NCBI non-redundant protein sequence databases with a rank-based evaluation similar as proposed by Podell & Gaasterland (2007), phylogenetic reconstructions based on HMMer 3 searches of all-frame translated end-sequences against the Pfam v. 25 database (Krause et al., 2008) and an evaluation of end-sequence tetranucleotide usage patterns (Teeling et al., 2004). These fosmids and the remaining 50 connecting fosmid candidates were sequenced at Genoscope (Évry Cedex, France) using the 454 FLX Ti platform and assembled using Newbler as described previously (Gómez-Pereira et al., 2010, 2012). The final dataset comprised 250 fosmids.

A.5.3 Fosmid re-assembly

Fosmid sequences were pooled by station and then re-assembled with SeqMan (Lasergene 8 software suite, DNAStar, Madison, WI, USA). The default setting was used. The assembly quality was checked via the program's strategy view option.

A.5.4 Taxonomic classification

Taxonomic affiliation of sequenced fosmids was done as described for end-sequences based on combined analysis for all genes of BLASTp hits to the NCBI non-redundant protein database and HMMer 3 hits to the Pfam v. 25 database (Supporting Information Table S4 - online).

A.5.5 Automated gene prediction and annotation

Gene prediction and annotation of all 226 contigs was done via the RAST server (Aziz et al., 2008). The RAST gene calls and annotations of the included 76 published fosmids differed only marginally from the published ones. Results for all contigs were downloaded and subsequently imported into a local installation of the GenDB (v. 2.2) annotation system (Meyer et al., 2003) for curation. CAZymes were annotated based on HMMer searches against the dbCAN database (Yin et al., 2012), BLASTp (Altschul et al., 1990) searches against the CAZy database (Cantarel et al., 2009, Lombard et al., 2014) and HMMer searches against the Pfam v. 28 database (Finn et al., 2010) using *E*-values derived from manual annotations of test data (Supporting Information Table S5 - online). CAZymes were only annotated when at least two of the three database searches yielded positive results. Peptidases were automatically annotated based on

batch BLASTp searches against the MEROPS 9.13 database (Rawlings et al., 2012) using the default *E*-value cutoff criterion of 10^{-4}. ABC transporter, TonB-dependent receptor, *susD* genes and sulfatase genes were automatically predicted based on HMMer 3 hits to the Pfam v. 28 database at $E \leq 10^{-5}$ using the following profiles: ABC_tran, ABC_membrane, ABC_membrane_2, ABC_membrane_3, ABC_tran_2, ABC2_membrane, ABC2_membrane_2, ABC2_membrane_3, ABC2_membrane_4, ABC2_membrane_5, ABC2_membrane_6, TonB_dep_Rec, SusD, SusD-like, SusD-like_2, SusD-like_3 and sulfatase. Annotated sequences were deposited at NCBI's Genbank (BioSample accessions SAMN04870880 and SAMN04870884).

A.5.6 Manual CAZyme annotation

CAZymes were identified based on homology with a selected subset of characterized enzymes from each CAZyme family. Initial annotations were validated by BLASTp searches against the UniProtKB/SWISSPROT database as of February 2014 (The UniProt Consortium, 2014) and HMMer searches against the Pfam v. 27 database (Punta et al., 2012). Each CAZyme was assigned to a CAZY family and, when possible, to an EC number. Abundant GHs from the multi-functional families GH3, GH5, GH13 and GH16 were subjected to an in-depth phylogenetic analysis to determine their substrate-specificities. Experimentally characterized proteins (Supporting Information Table S6 - online) were selected from the CAZy database for each activity within a given GH family and aligned to their contig homologs using MAFFT (FFT-NS-i iterative refinement method; BLOSUM62 amino acid substitution matrix) (Katoh & Standley, 2013). These alignments were used to calculate model tests and maximum likelihood trees with MEGA v. 6.0.6 (Kumar et al., 2004) with bootstrapping (100 resamplings). Annotation of ambiguous proteins was refined based on the proximity to characterized proteins in the phylogenetic trees.

Subcellular locations were predicted using CELLO v. 2.5 (Yu et al., 2006), PSORTb v. 3.0.2 (Yu et al., 2010) and HMMer searches against the TIGRfam profile (Selengut et al., 2007) TIGR04183 (Por secretion system C-terminal sorting domain). Only unambiguous consensus predictions were considered as reliable.

A.6 Acknowledgements

We thank the Captain and Crew of the FS Maria S. Merian for their support during cruise MSM03/01, J. Waldmann for bioinformatics and R. Hahnke for comparative PUL alignments. A. Gobet and G. Michel were supported by the National Research Agency of the French Government by the 'Blue Enzymes' ANR project (ANR-14-CE19-0020-01). This study was funded by the Max-Planck-Society, the German Science Foundation (DFG) and the FP6 EU program Network of Excellence Marine Genomics Europe.

Appendix B

Adaptive mechanisms that provide competitive advantages to marine bacteroidetes during microalgal blooms

Frank Unfried[1,2,3], Stefan Becker[2,4], Craig S. Robb[2,4], Jan-Hendrik Hehemann[2,4], Stephanie Markert[1,3], Stefan E. Heiden[1], Tjorven Hinzke[1,3], Dörte Becher[1,5], Greta Reintjes[2], Karen Krüger[2], Burak Avcı[2], Lennart Kappelmann[2], Richard L. Hahnke[6], Tanja Fischer[2], Jens Harder[2], Hanno Teeling[2], Bernhard Fuchs[2], Tristan Barbeyron[7,8], Rudolf I. Amann[2] and Thomas Schweder[1,3]

[1]Pharmaceutical Biotechnology, University Greifswald, Greifswald, Germany
[2]Max Planck Institute for Marine Microbiology, Bremen, German
[3]Institute of Marine Biotechnology, Greifswald, Germany
[4]MARUM, Center for Marine Environmental Sciences at the University of Bremen, Bremen, Germany
[5]Institute for Microbiology, University Greifswald, Greifswald, Germany
[6]DSMZ, Braunschweig, Germany
[7]National Center of Scientific Research/Pierre and Marie Curie University, Paris, France
[8]UMR 7139 Marine Plants and Biomolecules, Station Biologique de Roscoff, Roscoff, Bretagne, France

Published in *The ISME Journal* (DOI: 10.1038/s41396-018-0243-5).

Contributions to the manuscript:
Experimental concept and design: 0%
Acquisition of experimental data: 10%
Data analysis and interpretation: 15%
Preparation of figures and tables: 10%
Drafting of the manuscript: 0%

Electronic figure versions and supplementary material are available at https://doi.org/10.1038/s41396-018-0243-5.

B.1 Abstract

Polysaccharide degradation by heterotrophic microbes is a key process within Earth's carbon cycle. Here, we use environmental proteomics and metagenomics in combination with cultivation experiments and biochemical characterizations to investigate the molecular details of *in situ* polysaccharide degradation mechanisms during microalgal blooms. For this, we use laminarin as a model polysaccharide. Laminarin is a ubiquitous marine storage polymer of marine microalgae and is particularly abundant during phytoplankton blooms. In this study, we show that highly specialized bacterial strains of the *Bacteroidetes* phylum repeatedly reached high abundances during North Sea algal blooms and dominated laminarin turnover. These genomically streamlined bacteria of the genus *Formosa* have an expanded set of laminarin hydrolases and transporters that belonged to the most abundant proteins in the environmental samples. In vitro experiments with cultured isolates allowed us to determine the functions of *in situ* expressed key enzymes and to confirm their role in laminarin utilization. It is shown that laminarin consumption of *Formosa* spp. is paralleled by enhanced uptake of diatom-derived peptides. This study reveals that genome reduction, enzyme fusions, transporters, and enzyme expansion as well as a tight coupling of carbon and nitrogen metabolism provide the tools, which make *Formosa* spp. so competitive during microalgal blooms.

B.2 Introduction

Phytoplankton blooms produce large quantities of beta-glucans, such as laminarin, a soluble β-1,3-glucan with β-1,6 side chains. The breakdown of these polysaccharides by heterotrophic microbes is a central part of the marine carbon cycle. Diatoms alone are estimated to produce $\sim$ 5–15 Gt of laminarin per year as their storage compound, making it a major food resource for heterotrophic marine organisms (Alderkamp et al., 2007b). Bacterial laminarinase activities are abundant in ocean surface waters, but also within deeper parts of the water column and in sediments (Arnosti et al., 2005, Keith & Arnosti, 2001). This suggests laminarin-degrading bacteria and their laminarinases are common across the oceans. How bacteria compete for this abundant labile energy substrate is therefore of relevance for a better understanding of the marine carbon cycle. Although partially studied with model organisms in the laboratory (Kabisch et al., 2014, Labourel et al., 2014, 2015, Xing et al., 2015), the enzymes used for laminarin degradation by microbes in the wild remain largely unknown or uncharacterized.

For complete degradation of one polysaccharide, microbes must have an adapted glycolytic pathway that contains multiple enzymes, which individually address each of the

different glycosidic linkages and structural compositions present in the macromolecule. The genes of glycan-degrading pathways cluster in operons named polysaccharide utilization loci (PULs). Recent works suggest that each polysaccharide requires a corresponding PUL (for review, see Grondin et al. (2017)). Horizontal gene transfer, vertical inheritance, and gene loss distribute PULs asymmetrically among genomes of microbes, creating the molecular basis for polysaccharide resource partitioning (Hehemann et al., 2016, 2010, 2017). This might explain the occurrence of diverse bacterial communities in the human gut (Cockburn & Koropatkin, 2016, El Kaoutari et al., 2013, Martens et al., 2011, Ndeh et al., 2017) or in the oceans, whose members rely on different degradation products of the same polysaccharide to co-exist (Buchan et al., 2014, Needham & Fuhrman, 2016, Teeling et al., 2012, 2016). However, it remains unclear whether the degradation of complex carbohydrates is a community effort or mainly driven by highly specialized individual strains. Furthermore, how microbes effectively compete for the same polysaccharide resource, such as the abundant laminarin, is currently unknown.

In previous studies, we reported the high abundance (up to 24% of all bacteria) of the flavobacterial genus *Formosa* during diatom-dominated spring blooms off the North Sea island Helgoland (Teeling et al., 2012, 2016). Furthermore, high laminarin concentrations were measured at the same sampling site (Becker et al., 2017). Together, these findings suggested that *Formosa* spp. are prominent candidates for the recycling of laminarin during spring microalgae blooms.

In this study, we explored molecular strategies, which provide competitive advantages to the genus *Formosa* during microalgal blooms in general and for laminarin utilization in particular. We examined two strains, *Formosa* Hel3_A1_48 (referred to as strain A) and *Formosa* Hel1_33_131 (strain B), both of which were isolated from the same sampling location (Hahnke et al., 2015), and which are representative of two distinct taxonomical clades found during phytoplankton blooms (Chafee et al., 2017). The combination of high-resolution metaproteomics and metagenomics of spring bloom water samples with the detailed proteomic and biochemical characterization of the respective PUL in a cultured model strain (*Formosa* B) allowed us to show that a specialized enzyme repertoire represents one of the adaptive mechanisms that provide a competitive advantage in substrate exploitation. Using laminarin as a model substrate, we demonstrate how a microalgal glycan resource can promote the enrichment of individual dominating taxa from an initially diverse microbial community with similar metabolic functions. Our data indicate that *Formosa* B tightly couples glycan utilization with the uptake of nitrogen compounds. This suggests that a balanced carbon and nitrogen diet is required for competitive laminarin utilization during phytoplankton blooms.

B.3 Materials and methods

B.3.1 Growth experiments and physiological characterization

The investigated strains *Formosa* sp. Hel1_33_131 (*Formosa* strain B) and *Formosa* sp. Hel3_A1_48 (*Formosa* strain A) were isolated by dilution cultivation during a spring and a summer phytoplankton bloom, respectively, from surface water near the North Sea island Helgoland in the German Bight (Hahnke et al., 2015). Growth experiments were performed in a modified HaHa medium (Hahnke et al., 2015) (with 0.1 g L^{-1} peptone, 0.1 g L^{-1} casamino acids, 0.1 g L^{-1} yeast extract, 200 µM NH_4Cl, and 16 µM KH_2PO_4) with defined carbon sources as substrates at 12 °C during gentle shaking at 55 rpm. For the proteome analyses, described below, d-glucose and laminarin (L9634, Sigma-Aldrich Chemie GmbH, Taufkirchen, Germany) were used as carbon sources (concentrations: 2 g L^{-1}). In addition, the utilization of chitin (SAFSC9213, VWR) was tested in this medium and these cultures were used as a control condition for the in vitro proteome analyses with glucose and laminarin. All growth experiments were carried out in triplicates. Cells were harvested by centrifugation (15 min; 9500 × g; 4 °C), and the resulting pellets and supernatants were stored at -80 °C until use.

B.3.2 Genome sequencing, assembly, and annotation

For genome sequencing of the strains *Formosa* A (Hel3_A1_48) and B (Hel1_33_131) DNA was extracted according to the protocol of Zhou et al. (1996). Sequencing was performed at LGC Genomics (Berlin, Germany) using the 454 GS FLX Ti platform (454 Life Sciences, Branfort, CT, USA) using standard shotgun libraries. Draft genomes were assembled with Newbler v2.6 for Hel3_A1_48 from 640,093 reads (406,983,286 bp) and for Hel1_33_131 from 636,323 reads (410,253,204 bp), yielding 2,025,184 bp (77 contigs) and 2,727,763 bp (61 contigs), respectively. The remaining gaps were closed by PCR and Sanger sequencing, yielding circular assemblies of 2,016,454 bp for Hel3_A1_48 (*Formosa* A) and 2,735,158 bp for Hel1_33_131 (*Formosa* B). Gene prediction and annotation (including the phylogeny-guided carbohydrate-active enzyme (CAZyme) annotations provided in Supplementary Table S2 - online) were performed as described previously (Mann et al., 2013). Further bioinformatic analyses are described in Supplementary Information. Annotated genome sequences were submitted to NCBI's GenBank with the accession numbers CP017259.1 for *Formosa* sp. Hel3_A1_48 (*Formosa* strain A) and CP017260.1 for *Formosa* sp. Hel1_33_131 (*Formosa* strain B).

B.3.3 Proteome analyses

The soluble intracellular proteome, the enriched membrane-associated proteome, and the soluble extracellular proteome was characterized from exponentially growing cells of *Formosa* strain B. Details of the protein extraction and subproteome enrichment can be found in Supplementary Information.

Peptides were subjected to a reversed phase C18 column chromatography on a nano ACQUITY-UPLC (Waters Corporation, Milford, MA, USA) and separated as described by Otto et al. (2010). Mass spectrometry (MS) and MS/MS data were recorded using an online-coupled LTQ-Orbitrap Classic mass spectrometer (Thermo Fisher Scientific Inc., Waltham, MA, USA). We searched MS spectra against a target-decoy protein sequence database including sequences of *Formosa* B (Hel1_33_131) and of common laboratory contaminants.

Protein searches were performed using MaxQuant with the integrated Andromeda engine (Cox & Mann, 2008) with a peptide level FDR (false discovery rate) set to 0.01 (1%). Only proteins that could be detected in at least two out of three replicates were counted as identified. The automatically calculated iBAQ values (intensity-based absolute quantification; i.e., peak area divided by the sum of all theoretical peptides) were used to manually calculate riBAQ values (relative iBAQ; giving the relative protein abundance in % of all proteins in the same sample, Shin et al. (2013)) for semiquantitative comparisons between samples from different nutrient conditions. Tests for differential expression were performed using Perseus (Tyanova et al., 2016) v. 1.6.1.1 with Welchs t test (permutation-based FDR 0.05).

The mass spectrometry proteomics data are available through the ProteomeXchange Consortium (http://proteomecentral.proteomexchange.org) via the PRIDE partner repository (Vizcaíno et al., 2016) with the dataset identifier PXD007934.

B.3.4 Biochemical enzyme characterizations

The cloning of the FbGH17A gene (locus tag FORMB_24720) and the FaGH17B gene (locus tag FORMB_24740) is described in Supplementary Information. Cloning of the FbGH30 gene is described by Becker et al. (2017). Detailed information on the overexpression, enzyme refolding, and purification of these proteins can be found in Supplementary Information. For enzyme characterizations laminarin from *Laminaria digitata* (0.1% [w/v]; Sigma) was hydrolyzed over the course of 60 min at 37 °C with 100 nM purified enzyme ($\sim$ 5 µg mL^{-1} of FbGH30, FbGH17A, or FbGH17B) in 50 mM MOPS buffer at pH 7. The preparation and purification of debranched laminarin as well as

the determination of kinetic parameters of the three enzymes FbGH30, FbGH17A, and FbGH17B acting on native and debranched laminarin is explained in Supplementary Information. High performance anion exchange chromatography with pulsed amperometric detection (HPAEC-PAD) was applied for qualitative product analysis of the enzyme reactions (see Supplementary Information).

B.3.5 Protein crystallization and structure solution

Crystals of FbGH17A were obtained by hanging drop vapor diffusion of the protein with 12.3 mg mL^{-1} mixed 1:1 with a "well solution" (0.03 M $MgCl_2$, 0.1 M MOPS (pH 7), 9% PEG8K) supplemented with 15% ethylene glycol. The crystals were cryoprotected prior to freezing in the "well solution" supplemented with ethylene glycol to a final concentration of 30%. Crystals were frozen by flash freezing in liquid nitrogen in nylon loops. X-ray diffraction data were collected at the DESY P11 beamline. The structure was solved by molecular replacement using PHASER in the phenix suite (Adams et al., 2010, McCoy et al., 2007) using the pdb 4wtp (Qin et al., 2015). The model was built using BUCCANEER (Cowtan, 2006) and Coot, refined in REFMAC5 (Murshudov et al., 2011), and validated and deposited with pdb code 6FCG.

B.4 Results

B.4.1 Genome properties and phylogeny

We sequenced and annotated the genomes of the *Formosa* strains A and B. Both have a single chromosome with a GC content of 36.4 and 36.6%, respectively. With 2,016,454 bp (strain A) and 2,735,158 bp (strain B) they possess small genomes compared with other marine polysaccharide-degrading *Flavobacteriia* (Barbeyron et al., 2016b, Bauer et al., 2006, Kabisch et al., 2014, Mann et al., 2013). Strain A has 1913 predicted genes including 1866 coding sequences (CDS), 40 tRNAs genes, and 2 rRNA operons (identical 5 S, 16 S, 23 S rRNA genes), whereas the strain B genome encodes 2675 predicted genes with 2628 CDS, 39 tRNAs genes, and 2 rRNA operons (identical 5 S, 16 S, 23 S rRNA genes).

Phylogenetic analyses based on 16 S rRNA gene sequences indicate that the *Formosa* strains A and B are representatives of two previously uncultured clades of the genus *Formosa*. These occur not only in the North Sea, but also in surface waters from coastal and open ocean sites throughout the world (Supplementary Figure S1 - online). Of the

33 full-length *Formosa* 16 S rRNA sequences obtained from 2009 spring bloom bacterioplankton (Teeling et al., 2012), 16 were > 99% identical to *Formosa* sp. Hel1_33_131 (*Formosa* B) (Supplementary Figure S1 - online).

B.4.2 *Formosa* genomes encode PULs for laminarin degradation

Genome annotation suggested that the *Formosa* strains A and B are specialized polysaccharide degraders, which concentrate their genetic potential on a small set of sugars. *Formosa* A contains seven PULs (Supplementary Figure S2 - online) and 28 glycoside hydrolases (Supplementary Table S1 - online), whereas *Formosa* B contains six PULs (Supplementary Figure S3 - online) and 21 glycoside hydrolases (Supplementary Table S1 - online). This is a very small repertoire, even compared with other marine *Bacteroidetes* isolated from algal blooms (Xing et al., 2015). These small CAZyme repertoires contrast particularly with those of generalist polysaccharide degraders isolated from macroalgae, such as *Formosa* agariphila (Mann et al., 2013), which have broad polysaccharide-degrading capacity. *F. agariphila* has, for example, a genome size of 4.48 Mbp, and 84 glycoside hydrolases in 13 PULs (Supplementary Table S1 - online).

To functionally characterize laminarin-specific PULs of the *Formosa* strains A and B, we searched the genomes for enzymes belonging to known laminarinase-containing families (Supplementary Table S1 - online). We found putative laminarinases of the families GH16 and GH17 but also enzymes of the GH3 and GH30 families as well as a member of the newly described GH149 family (Kuhaudomlarp et al., 2018), located in close proximity to TonB-dependent receptors (TBDR) and SusD-like proteins, which are indicators of PULs (Sonnenburg et al., 2010, Tang et al., 2012). Our results suggest that there are three putative laminarin-specific genomic PULs in both *Formosa* strains (Supplementary Figure S2–S5 - online).

The laminarin PULs 1 and 2 of *Formosa* A and B revealed a high synteny with PULs from other bacteroidetal strains (Supplementary Figure S4 - online) from North Sea surface water (Bauer et al., 2006, Hahnke et al., 2015, Panschin et al., 2016). This points to a potential for competition between those groups, but also suggests that this part of the laminarin utilization machinery is highly conserved. However, the *Formosa* B PULs 1 and 2 are enlarged with laminarinases and transporters that are partially not present in the other bacteria. Moreover, the entire PUL 3 of *Formosa* B is missing in these other strains (Supplementary Figure S5 - online). Instead, *Formosa* B's PUL 3 shows synteny to PULs of other marine *Flavobacteriia*, which do, however, not possess the PULs 1 and 2 (Supplementary Figure S5 - online).

B.4.3 Laminarin elicits the expression of specific polysaccharide utilization loci in *Formosa*

Incubation experiments with fluorescently labeled (FLA) laminarin revealed the ability of *Formosa* B to quickly react and take up laminarin. *Formosa* B accumulated high amounts of FLA-laminarin after just 5 min of incubation (Fig. 1a). Additionally, the halo-like staining pattern showed that the FLA-laminarin was imported into the periplasm of the cells by a "selfish" uptake mechanism (Cuskin et al., 2015, Reintjes et al., 2017). Selfish substrate uptake is dependent on the presence of SusCD-like transporters and secures an enrichment of substrate in the periplasmic space without diffusive loss (Reintjes et al., 2017).

To elucidate the metabolism of *Formosa* B on laminarin and to verify whether laminarin specifically controls the expression of the genomically detectable PULs we performed cultivation experiments with this bacterium with purified laminarin as growth substrate. Growth curves of *Formosa* B in HaHa medium with laminarin, glucose and only protein extracts, respectively, are shown in Fig. 2.1b. We used proteomics to record the global protein expression patterns with these substrates. We investigated (i) the soluble intracellular proteome, (ii) the enriched membrane proteome, and (iii) the extracellular proteome (see Supplementary Information and Supplementary Tables S2A–C - online). These comparative analyses showed that although glucose is the monomer of laminarin, the utilization of either carbon source led to quite different proteomic signatures in different functional protein categories, such as in nucleotide, lipid, and coenzyme metabolism as well as in carbohydrate metabolism and transport (Supplementary Figure S6 - online). About 100 proteins were significantly higher abundant or only found in laminarin incubations in *Formosa* B (Supplementary Figure S7 - online, Supplementary Table S3 - online). Of all three substrates, laminarin elicited the strongest expression of the three laminarin PULs of *Formosa* B (Fig. 2.1c), which is indicative of specific and tightly controlled expression. The SusD-like protein (FORMB_10080) of PUL 1 and the GH16 (FORMB_m24690) of PUL 2 were exclusively expressed with laminarin but not with the other substrates. Furthermore, the expression of PUL 3 was exclusively induced by laminarin and not detectable with glucose or only peptone (Fig. 2.1c). The specific response of the Formosa PULs to laminarin and not to glucose implies that the three-dimensional structure of laminarin might be the key to induce the expression of these PULs.

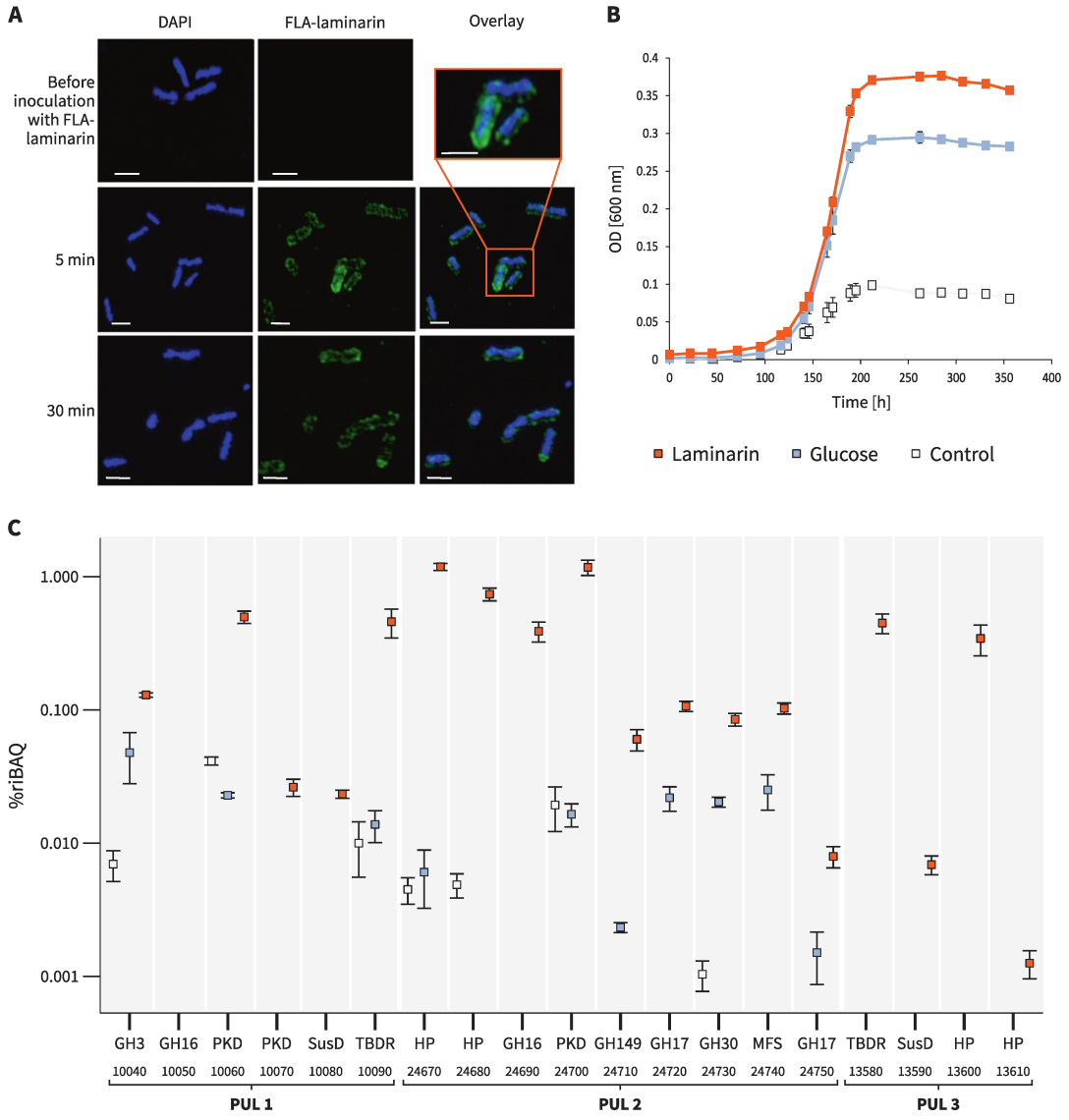


Figure 2.1: Laminarin utilization of *Formosa* B. **a** SR-SIM of *Formosa* B cells before inoculation with FLA-laminarin at 5 and 30 min after incubation with FLA-laminarin. Images show cell staining by DAPI (left, blue), FLA-laminarin (middle, green), and an overlay showing both FLA-laminarin staining and DAPI (right). Scale bar = 1 µm **b** Growth curves of three biological replicates at 12 °C in modified HaHa_100V medium (Hahnke et al., 2015) with 2 g L^{-1} laminarin or 2 g L^{-1} glucose. The "control" culture contained only 0.1 g L^{-1} peptone, 0.1 g L^{-1} yeast extract, and 0.1 g L^{-1} casamino acids but no additional carbon sources. **c** Expression profile and gene organization of the laminarin utilization PULs 1–3 in *Formosa* B. Relative protein abundances (in %riBAQ) of PUL-encoded proteins detected in the membrane protein fractions of each three independent cultures grown on laminarin (orange), glucose (blue), and chitin (control, gray) are shown (for riBAQ values see Supplementary Tables S2A and S3 - online). Putative protein functions (e.g., GH3) and the respective locus tags (e.g., 10040) are indicated. The squares represent the mean values of the replicates for every protein and each substrate. The error bars refer to the standard error of the mean. Proteins that could be detected in at least two out of three independent biological replicates of each substrate condition are shown (for individual replicate numbers see Supplementary Table S2A - online). GH, glycoside hydrolase; PKD, PKD-domain containing protein; SusD, SusD-family protein; HP, hypothetical proteins; MFS, major facilitator superfamily; TBDR, TonB-dependent receptor.

B.4.4 Biochemical analysis of laminarinases expressed by *Formosa* spp

To functionally characterize PUL-encoded proteins and to map the laminarin degradation pathway, we cloned and biochemically analyzed putative laminarinases the function of which could not be merrily solved by comparative sequence analyses with known enzyme functions. We cloned and examined the genes encoding FbGH17A (locus tag: FORMB_24720), FaGH17B (locus tag: FORMB_24740), and FbGH30 (locus tag: FORMB_24730). As all three proteins are encoded in a single gene cluster, we hypothesized that these enzymes might work together in spatial proximity. To test this hypothesis, we conducted a series of biochemical experiments, which revealed that the FbGH30 enzyme hydrolyzed the β-1,6-linked glucose side chains of laminarin (K_M: 3.1 ± 0.2 mM and K_{cat}/K_M: 21124 $M^{-1}s^{-1}$) (Fig. 2.2a), whereas it was inactive on the debranched substrate. The enzyme FbGH17A hydrolyzed both the debranched laminarin product of FbGH30 and the native laminarin, although with a markedly higher specific activity on the debranched product (K_M: 1.6 ± 0.1 mM; K_{cat}/K_M: 36056 $M^{-1}s^{-1}$) than on laminarin itself (K_M: 4.3 ± 0.1 mM; K_{cat}/K_M: 25744 $M^{-1}s^{-1}$) (Fig.2.2b). Preference for debranched laminarin was even more pronounced with FbGH17B, which only hydrolyzed debranched laminarin (K_M: 2.6 ± 0.3 mM; K_{cat}/K_M: 30803 $M^{-1}s^{-1}$) (Fig.2.2c) and was inactive on the branched form.

To elucidate how these enzymes work together in successive laminarin degradation, we used high-performance liquid chromatography method with photo diode array detection analyses. The data indicated an enzymatic functional cascade in three steps (Fig.2.2d): The exo-acting β-1,6-glucosidase FbGH30 removes the glucose side chains from laminarin (Supplementary Figure S8A - online). The endo-acting β-1,3-glucan hydrolase FbGH17A degrades the remaining debranched laminarin into oligosaccharides (Supplementary Figure S8B - online). The exo-acting β-1,3-glucosidase FbGH17b processes these oligosaccharides into glucose (Supplementary Figure S8C - online). FbGH17b is part of a multimodular protein, which is encoded by a gene that also codes for an N-terminal major facilitator superfamily (MFS) transporter, suggesting that hydrolysis and product uptake might be coupled. The MFS transporter contains 12 transmembrane-spanning helices (as predicted by Phyre2 (http://www.sbg.bio.ic.ac.uk/phyre2/html/page.cgi?id=index)), with the last C-terminal helix and the attached GH17 domain (Fig. 2.2e), which would enable the simultaneous cleavage of oligosaccharides and the sugar transported through the MFS. Blast analysis revealed that this fusion is common among marine *Flavobacteriia*, suggesting that such multi-modular transporter-associated enzyme may be a conserved mechanism for boosting laminarin utilization.

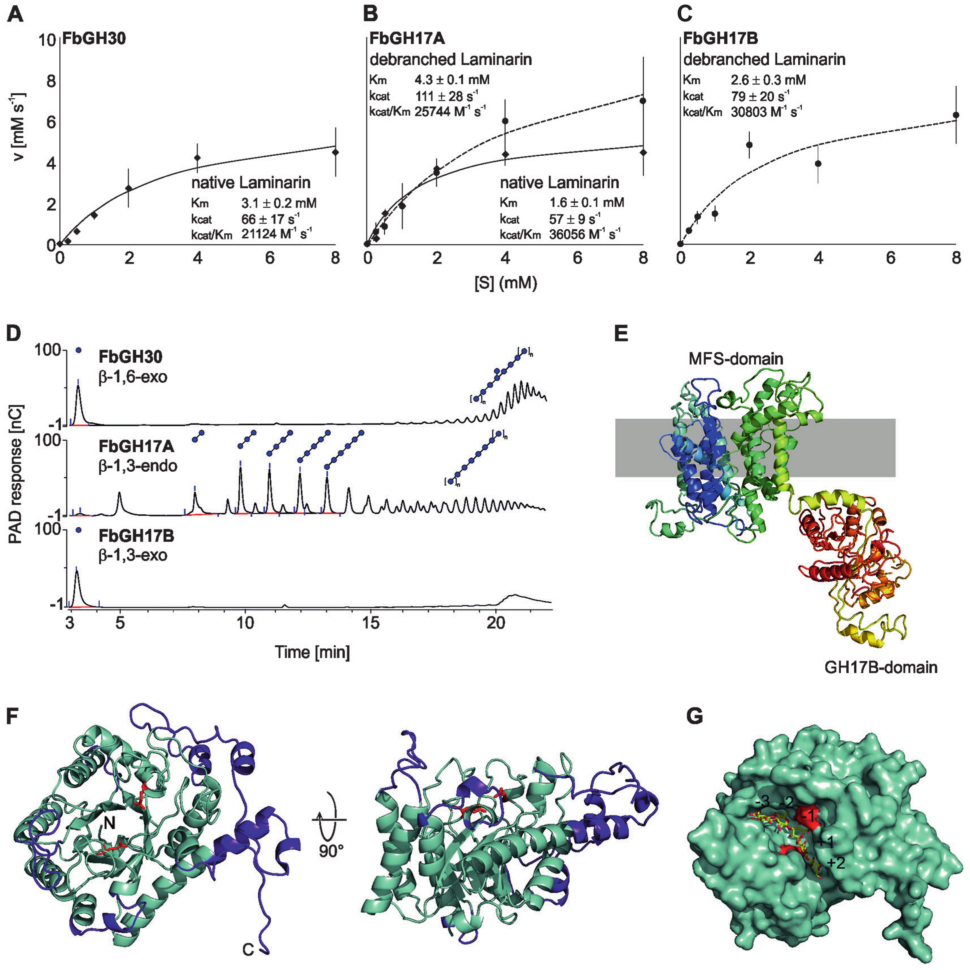


Figure 2.2: Biochemical characterization of different laminarinases from *Formosa* B. Michaelis–Menten Kinetic (**a**) of FbGH30 on native laminarin, **b** of FbGH17A on native and debranched laminarin and c of FbGH17B on debranched laminarin (native laminarin is illustrated by the solid lines and debranched laminarin by the dashed lines). **d** Visualization of all three enzymatic activities was done using HPAEC-PAD. FbGH30 hydrolyzed native laminarin. After this debranching reaction, the laminarin was purified to remove glucose for the following steps. This debranched laminarin was used in the FbGH17A reaction. FbGH17B hydrolyzed the products of the previous FbGH17A reaction without any further purification in between. **e** 3D structure model of both the MFS-domain and the associated FbGH17B-domain and its potential arrangement within the inner-membrane. The modeling was performed using Phyre2. **f** The overall structure of FbGH17A is displayed in cyan with the additions colored purple. Highlighted in red are the catalytic residues. The N- and C-termini are labeled. **g** A surface view of FbGH17A with a modeled substrate complex shown as sticks in yellow from a GH17 transferase of *Rhizomucor miehei* with laminaritriose and laminaribiose in the -3 to -1 and +1 to +2 subsites, respectively.

In order to examine the molecular basis of substrate specificity the X-ray crystal structure of GH17A was solved (Fig. 2.2f, see Supplementary Information). Compared to the monomeric GH17 structures, FbGH17A has significant insertions and is larger. The structure of GH17A allows for the deduction of the molecular basis of substrate specificity for the laminarinase. Based on the GH17 complexes obtained for *R. miehei* (Qin

et al., 2015), a model was generated of a laminarin product involving five monomers bound to the catalytic groove, two on the aglycone side, and three on the glycone side (Fig. 2.2g). The reducing and non-reducing ends of the modeled glycan are free, suggesting the protein can act in the middle of the chain as expected for an endo-acting glycoside hydrolase (see Supplementary Information). Furthermore, given the conformation of the modeled glycan, 6-O-β-glucose branching would be possible only at subsite +1 and anything further away (+3 or -4). In other words, within the native polysaccharide the enzyme would need a stretch of at least three free β1,3-glucose moieties to act. This structural data supports the observation that GH17 activity on laminarin is bolstered by the action of the debranching enzyme GH30.

B.4.5 Laminarin stimulates the co-expression of selected peptidases and transporters

Formosa B encodes 69 peptidases in its relatively small genome (2.7 MB). Other laminarin-degrading marine *Bacteroidetes* like *Gramella forsetii* KT0803T (79 peptidases), *Polaribacter* sp. Hel1_85 (84 peptidases), *Jejuia pallidilutea* (58 peptidases), and *Flaviramulus ichthyoenteri* (63 peptidases) show a comparable number of peptidase genes, although their genomes are around twice as large as that of *Formosa* B. Our proteome analysis of *Formosa* B revealed that 41 peptidases are expressed in the presence of laminarin (Supplementary Table S4 - online). Nine of these peptidases showed a significantly higher protein abundance on laminarin in the enriched membrane proteome, compared with glucose or the control culture, or were exclusively found after incubation with laminarin (Fig. 2.3a and Supplementary Table S3 - online). In addition, a putative peptide ABC transporter ATP-binding protein (FORMB_10920) and a putative oligopeptide permease ABC transporter protein (OppC; FORMB_20460) were detected, which showed a significantly higher abundance under laminarin conditions (Supplementary Table S3 - online). This indicates a coupling of the peptide metabolism with laminarin utilization in *Formosa* B.

An exceptionally high expression with glucose and laminarin was visible for a putative porin (FORMB_11920, 10% riBAQ; Fig. 2.3b, c), an outer membrane protein, which was not detectable in the control cultivations with peptone (Supplementary Table S3 - online). The porin-encoding gene is located in an operon with a putative ammonium transporter and clusters with several genes involved in nitrogen metabolism, including two putative glutamate synthase genes and an additional supposed ammonium transporter (Fig. 2.3c). All nitrogen metabolism-related genes in the direct vicinity of the porin-encoding gene were only found to be expressed with glucose and laminarin in

peptone-containing cultures in comparison with the peptone-only control culture without these carbon sources (Fig. 2.3c).

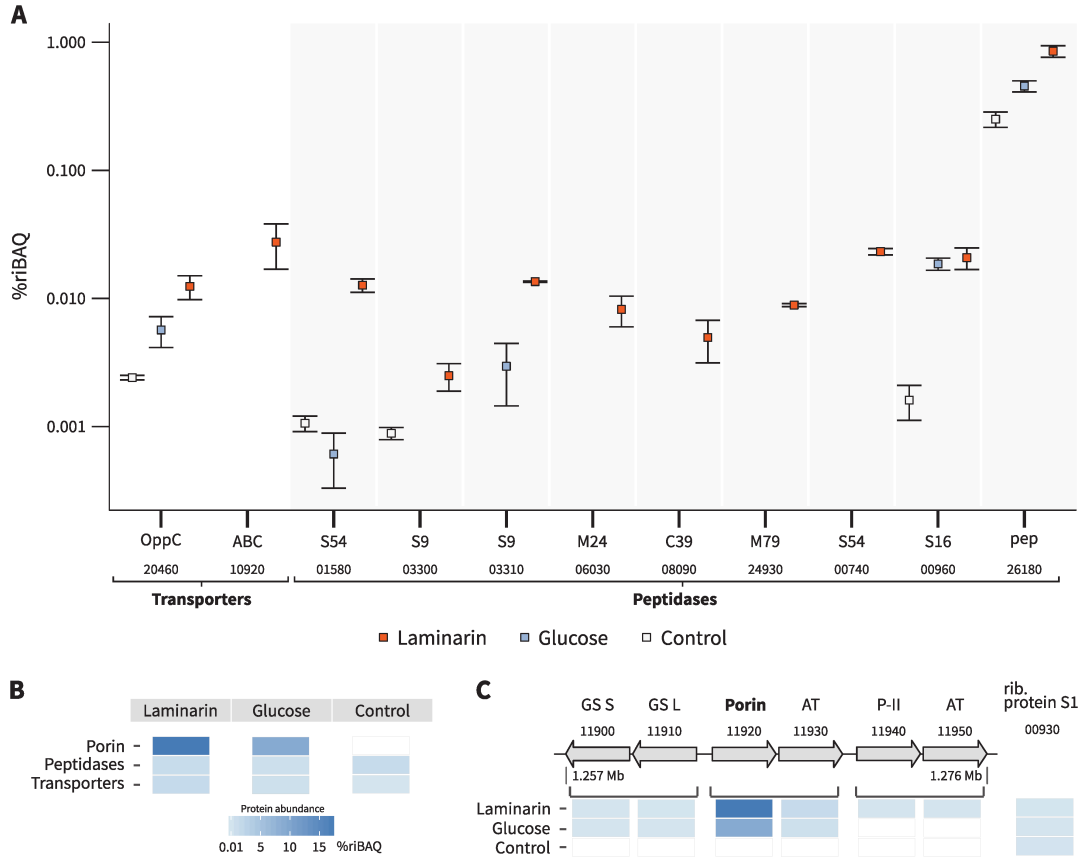


Figure 2.3: Abundance of peptidases and putative peptide transporter proteins of *Formosa* B. **a** Proteomic signatures of selected proteases and two putative peptide transporters, which showed an increased expression under laminarin conditions. The two laminarin-induced peptidases shown on the right (00960, 26180) were also detected in the metaproteome of a spring bloom in 2009 (see Supplementary Table S8A - online). Relative protein abundances are depicted as % riBAQ values. The putative peptidase families and the respective locus tags are indicated. The squares represent the mean values of the three replicates for every protein and each substrate. The error bars refer to the standard error of the mean. Proteins that could be detected in at least two out of three independent biological replicates of each substrate condition are shown (for individual replicate numbers see Supplementary Table S2A - online). **b** Comparison of total protein abundances (in %riBAQ) of peptidases (see Supplementary Table S4 - online), nitrogen-associated transporters and the porin (FORMB_11920) in *Formosa* B under the three investigated substrate conditions (see Supplementary Table S3 - online). **c** Genomic structure of the porin-encoding cluster and the abundance patterns of the corresponding proteins. Brackets indicate putative operons. Protein functions and the respective locus tags are indicated. GS S: glutamate synthase subunit S, GS L: glutamate synthase subunit L, AT: ammonium transporter, P-II: nitrogen regulatory protein P-II.

B.4.6 In situ abundance and relevance of *Formosa* strain A and B

We investigated the in situ abundance of the *Formosa* strains A and B by recruiting *Formosa* reads from the 44 metagenomes of the years 2009–2012 from Helgoland bacterioplankton samples (Teeling et al., 2016). At the ≥ 95% average nucleotide identity (ANI) threshold, the strain A and B genomes recruited up to 0.28% and 2.94% of individual metagenomic reads in 2009, 0.04% and 0.99% for 2010, 0.03% and 0.98% for 2011, and 0.02% and 0.21% for 2012 (Supplementary Table S5 - online), respectively. The mapped reads covered up to 91%, 99%, 97%, and 94% of the strain B genome from 2009 to 2012, respectively, and only up to 58% of the strain A genome in 2012 (Fig. 2.4 and Supplementary Table S5 - online). This suggests that strain B was recurrent and abundant during the spring bloom events, whereas strain A was likely more representative for late summer blooms reaching highest abundances of mapped reads in September 2009. Reads mapped to the *Formosa* strain B genome with 70–93% ANI suggest the presence of other closely related *Formosa* spp. during the spring blooms of 2009 to 2012 that reached up to 6.84%, 1.2%, 5.1%, and 1.8% of the metagenome reads, respectively (Fig. 2.4). Altogether, these results indicate that strain B is one of the representatives of the recurrent *Formosa* clade during North Sea spring microalgae blooms (Teeling et al., 2012, 2016).

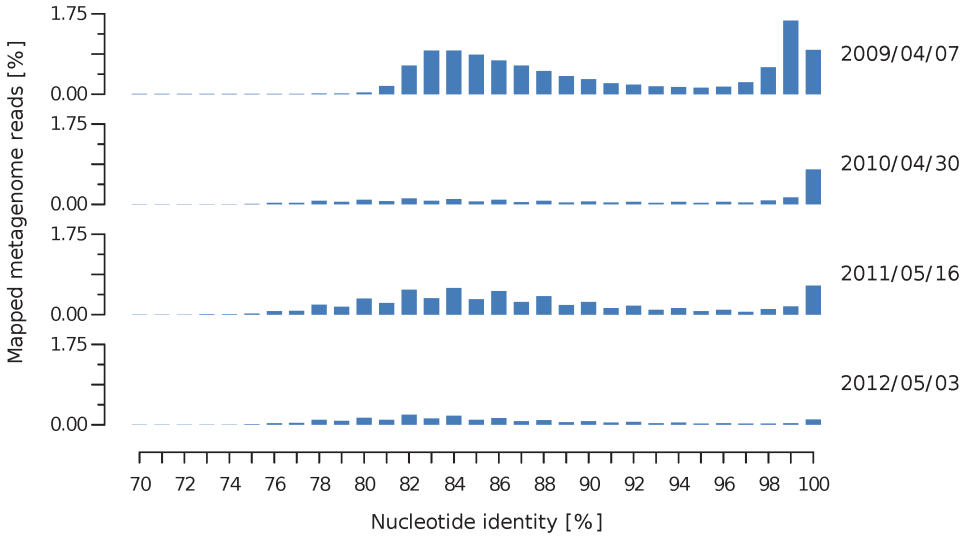


Figure 2.4: Relative abundance of *Formosa* strain B and related species during four spring phytoplankton blooms indicated by the percentage of metagenomics reads mapped at different nucleotide identities. Reads recruited at ≤ 93% nucleotide identity represent other *Formosa* spp. that are abundant during the spring bloom events at Helgoland, Germany from 2009 until 2012. Dates on the right indicate the four metagenomes (i.e., time points), which produced highest mapping coverage with the *Formosa* strain B genome in their respective years. For a summary of all 44 metagenomes (up to 18 time points per year) and their mapping results see Supplementary Table S5 - online.

B.4.7 Identification of *Formosa*-specific enzymes and transporters during microalgal blooms

All three *Formosa* B PULs were completely covered by metagenomic contigs of the spring bloom in 2009 and 2010 (Fig. 2.5a and Supplementary Tables S6–S7 - online), and partially covered in the metagenomes of 2011 and 2012 (Supplementary Table S7 - online). This illustrates the strong selection pressure imposed by laminarin on this pathway during four consecutive annual spring phytoplankton blooms in the North Sea.

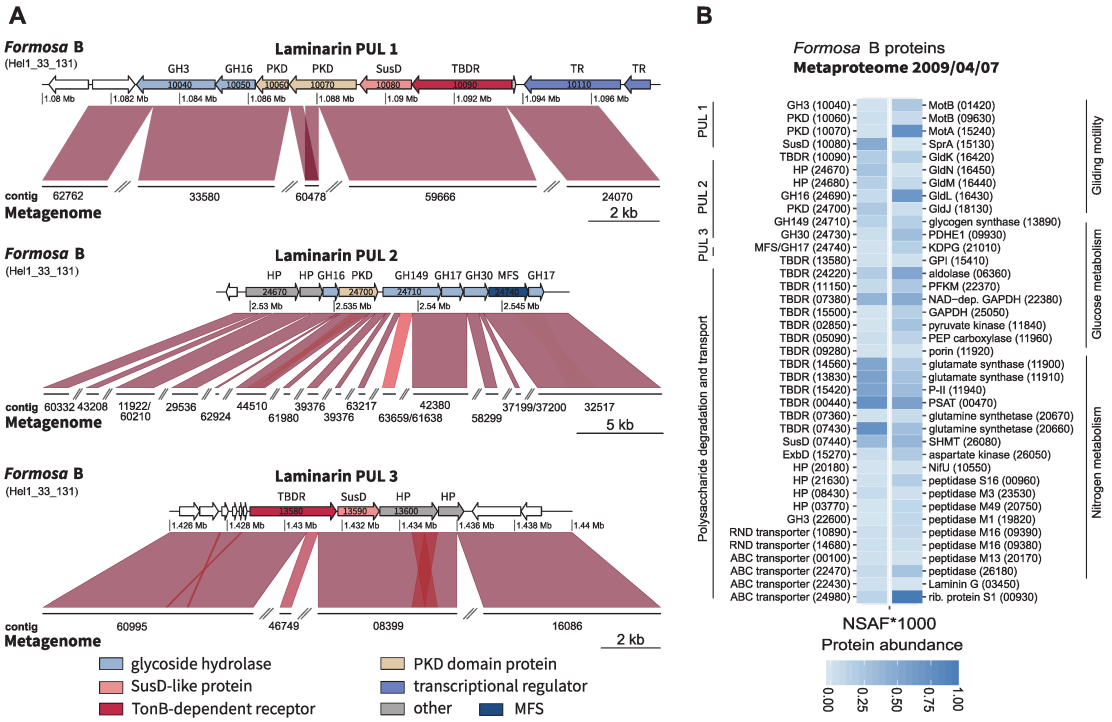


Figure 2.5: Detection of *Formosa* strain B laminarin PULs in the Helgoland spring bloom metagenome and metaproteome in 2009 (Teeling et al., 2012). **a** Synteny between the laminarin PULs of *Formosa* sp. Hel1_33_131 and partial PUL sequences in the metagenomes from 2009/04/07. The sequence comparisons were performed with Bl2seq (BLASTn, E value 1e-5). Sequence similarities are depicted by red hues for direct comparisons. Darker colors correspond to higher identities. Gene locus tags are subsequent numbers within PULs and are indicated in the figure for most of the genes (for visibility's sake, gene names of very small genes were omitted). **b** Heatmap of the relative abundance of *Formosa* strain B proteins (displayed as normalized spectral abundance factor values, NSAF*1000) detected in the metaproteome from 07 April 2009. Displayed are selected proteins, which likely play a role in polysaccharide or protein utilization. A highly abundant ribosomal protein of *Formosa* strain B (rib. protein S1, lower right) is also displayed as a reference to illustrate the high abundance of polysaccharide utilization-specific proteins during the bloom condition. Gene locus tag numbers are given in parenthesis. GH, glycoside hydrolase; PKD, PKD-domain containing protein; SusD, SusD-family protein; TBDR, TonB-dependent receptor; HP, hypothetical protein; MFS, major facilitator superfamily; ExbD, subunit of the Ton system for energy transduction; MotB, motor rotation protein; Gld, gliding motility proteins; PDHE, pyruvate dehydrogenase E1 component; KDPG, 2-dehydro-3-deoxyphosphogluconate aldolase; GPI, glucose-6-phosphate isomerase; PKFM, 6-phosphofructokinase; GAPDH, glyceraldehyde 3-phosphate dehydrogenase; P-II, nitrogen regulatory protein P-II; PSAT, phosphoserine aminotransferase; SHMT, serine hydroxymethyltransferase.

We examined the presence of polysaccharide degradation- and consumption-related proteins of the *Formosa* strains A and B in the *in situ* metaproteomes of spring blooms in 2009 and 2010 (Supplementary Table S8 - online). The proteome analysis of the planktonic bacterial fraction sampled during the spring bloom on 7 April 2009 uncovered 46 proteins from *Formosa* strain A and 361 proteins from *Formosa* strain B. Remarkably, several marker proteins from the putative laminarin-specific *Formosa* B PULs were highly abundant (Fig. 2.5b and Supplementary Table S8A - online) in the metaproteome samples. This analysis identified 13 proteins of the PULs 1, 2, and 3 (see also Supporting Information) and thus indicated that a significant proportion of *Formosa* B's laminarin PULs were expressed in situ during the spring bloom in 2009. Although the metaproteome analysis of 2010 uncovered fewer proteins from both *Formosa* strains, three marker proteins of PUL 1 from *Formosa* B were detected in the environmental samples (see Supplementary Information and Supplementary Table S8B - online).

Besides glycoside hydrolases and laminarin-specific transporter proteins, we also identified several *Formosa* B proteins in the environmental metaproteome samples of 2009, which are involved in the central catabolism of the monosaccharide glucose, the product of laminarin hydrolysis (see Fig. 2.5b, Supplementary Information and Supplementary Table S8A - online). This includes nearly all glycolytic enzymes as well as a putative glycogen synthase of *Formosa* B. These data indicate that the *Formosa* B strain substantially contributed to laminarin degradation and turnover during a diatom-driven phytoplankton bloom.

In addition, several proteins of *Formosa* B involved in nitrogen metabolism could be detected in the metaproteome analyses of the spring bloom 2009 (Supplementary Table S8A - online). This includes the putative porin (FORMB_11920), a peptide ABC transporter ATP-binding protein (FORMB_10920), an oligopeptide permease ABC transporter protein (OppC; FORMB_20460), and eight peptidases (Fig. 2.5b). This underlines a strong coupling of the peptide metabolism with laminarin utilization of *Formosa* B under in situ conditions.

B.5 Discussion

This study provides detailed insights in the adaptations which make *Formosa* strains successful competitors in the early breakdown of organic matter during diatom blooms. Combining comparative *in vitro* and *in situ* proteogenomics with biochemical enzyme characterization reveals that the key to this process is the sensing and utilization of laminarin. Our data indicate that this polysaccharide is used in two ways: as a major

source of energy, and as a signal molecule, which induces transporters and digestive enzymes to use also other compounds released from the lysis of diatom cells.

The two environmentally relevant *Formosa* strains examined in this study feature streamlined genomes, which are significantly smaller than those of many other marine *Flavobacteriia*. With a lower number of total proteins to synthesize, *Formosa* A and B can dedicate a higher relative proportion of their genomic and proteomic resources to the digestion of laminarin. Their CAZyme repertoire is strongly reduced compared to versatile polysaccharide degraders such as *F. agariphila* (Mann et al., 2013) and *Zobellia galactanivorans* (Barbeyron et al., 2016b), which were isolated from macroalgae. It is, however, similar to another member of North Sea spring bacterioplankton, *Polaribacter* sp. Hel1_33_49 (Xing et al., 2015). In contrast to macroalgae-associated laminarin-degrading bacteria, such as *Z. galactanivorans* (Groisillier et al., 2015), neither of the *Formosa* strains possesses a mannitol dehydrogenase, which indicates a specialization of *Formosa* A and B to chrysolaminarin. This type of laminarin lacks mannitol residues and is preferentially produced by diatoms.

We found a specific laminarin protein abundance pattern in *Formosa* B, which differs from the protein expression pattern in presence of the sugar monomer of this polysaccharide, glucose. A similar laminarin-specific control of gene expression was suggested for the marine flavobacterium *G. forsetii* (Kabisch et al., 2014). Interestingly, this laminarin-specific proteome signature of *Formosa* B includes not only the proteins required for laminarin uptake and utilization, but also peptidases and transporters for amino-acid utilization. The *Formosa* cells, upon sensing of laminarin, thus appear to react in two ways: First, they enhance the expression of outer membrane proteins to degrade and rapidly transport the energy molecule laminarin into their periplasm, utilizing the selfish polysaccharide uptake mechanism recently demonstrated for marine *Flavobacteriia* (Reintjes et al., 2017). Second, the abundance of amino-acid and nitrogen metabolism-related proteins is increased to boost the recycling of nitrogen building blocks, which are required for rapid growth of *Formosa* bacteria and become available simultaneously with laminarin upon algal lysis.

Formosa strain B possesses an extended repertoire of laminarin-specific enzymes and transporters, which is larger than that of other laminarin-degrading bacteria such as *Polaribacter* sp. Hel1_33_49 (Xing et al., 2015) or *G. forsetii* (Kabisch et al., 2014). Our subproteome and bioinformatic analyses indicate that many laminarin-degrading enzymes of *Formosa* B are surface-tethered or localized in the periplasmic space and in the cytoplasmic membrane, respectively. The different TBDR, laminarinases, transporters, and additional enzymes combine complementary activities into an efficient laminarin disassembly line for degradation and uptake (Fig. 2.6). The biochemical experiments

presented here support the annotation of the conserved cluster of genes as encoding for a laminarin utilization pathway. Here, two enzymes that are likely residents of the periplasm are shown to work together towards the complete degradation of laminarin in a highly specific manner. The X-ray crystal structure of GH17A reveals the possible molecular determinants of substrate specificity and the propensity of the enzyme to be more active on unbranched laminarin.

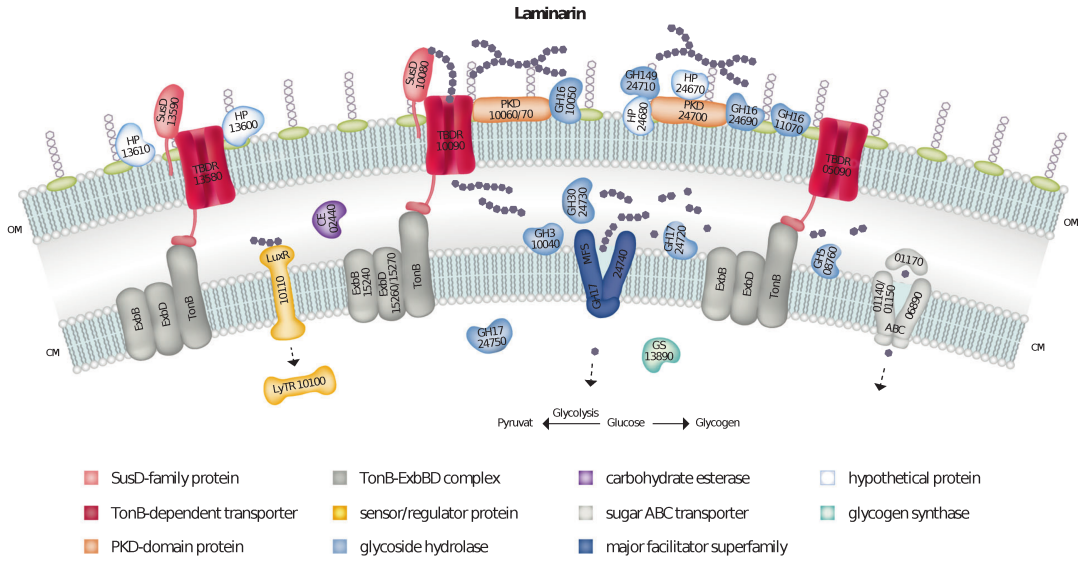


Figure 2.6: A tentative model of laminarin utilization pathways in *Formosa* B. Protein localizations were predicted in silico according to Romine (2011) and were deduced by subproteome analyses (see also Supplementary Table S3 - online). Additional biochemical experiments are required to ascertain this model.

The multi-modular protein FORMB_24740 (FbGH17b) combines a glycoside hydrolase (GH17) with a membrane spanning transport protein and may represent an adaptive mechanism for laminarin utilization. The integration of the transport and hydrolysis processes into a single protein could facilitate improved consumption of the sugars by increasing the activity of the fused GH17. Increased activity would also reduce the necessary enzyme copy number, and thereby resource consumption for synthesis of this protein. To our knowledge, such a transporter-CAZyme-fusion has not been described for other bacteria as yet, but its conservation in nature suggests that this could provide a significant benefit.

The exceptionally strong accumulation of a putative porin in glucose- and laminarin-controlled cultures and the co-induction of the 12 surrounding genes of this porin-encoding genomic cluster, all of which play a role in nitrogen metabolism, could indicate a function of this transporter protein in the uptake of peptides as nitrogen and amino-acid source. Endo-acting proteases might degrade proteins released by lysed microalgae

into peptides, which are then imported through the porin into the periplasm. An efficient capture of these peptides with a highly abundant porin system might be especially useful in the highly diffusive marine environment. With laminarin and glucose as easily metabolizable carbon sources, such a strategy could be crucial for a balanced carbon and nitrogen diet.

Members of the phylum *Bacteroidetes* are primary degraders of microalgal polysaccharides during phytoplankton blooms, and are therefore key players in marine carbon cycling. However, underlying enzymatic mechanisms and adaptations that drive the specialization of these highly competitive bacteria remain obscure. We reveal and prove in this study the specific activity and ecological niche of two abundant marine *Bacteroidetes* strains in complex microbial communities during diatom-driven phytoplankton blooms. Our results show an extraordinary degree of specialization for the *Formosa* strains A and B, which enable these marine *Bacteroidetes* to successfully compete for laminarin against a multitude of other laminarin-degrading microbes in bloom situations (Alderkamp et al., 2007b, Bennke et al., 2016, Cardman et al., 2014, Kabisch et al., 2014, Xing et al., 2015). Our data furthermore indicate that fast growth on beta-glucans such as laminarin requires a balanced diet that also includes nitrogen sources like peptides. The induction of several cell wall-associated peptidases and peptide-specific transporters in *Formosa* B during growth on laminarin suggests that these bacteria pursue a complex uptake strategy, which encompasses both sugars and nitrogen compounds. This may make marine *Flavobacteriia* so successful in their diffusion-open environment.

B.6 Acknowledgements

We are grateful to Jana Matulla for technical assistance, Sebastian Grund for mass spectrometry analyses, Carol Arnosti and her laboratory for producing the FLA-labeled laminarin, and Nicolas Terrapon for critical comments on the putative function of selected PUL-encoded proteins. We thank the crew of the Biological Institute Helgoland of the Alfred-Wegener-Institute, especially Antje Wichels, for their excellent logistic support during our sampling campaigns at the Kabeltonne Helgoland from 2009 until 2012. The Federal Ministry of Education and Research (BMBF) funded parts of this work within the “Microbial Interactions in Marine Systems” project (MIMAS project 03F0480A). The work was also financially supported by the DFG in the framework of the research unit FOR2406 “Proteogenomics of Marine Polysaccharide Utilization” (POMPU) by grants of R. Amann (AM 73/9-1), H. Teeling (TE 813/2-1), B. Fuchs (FU 627/2-1), D. Becher (BE 3869/4-1), J.-H. Hehemann (HE 7217/2-1), and T. Schweder (SCHW 595/10-1). F. Unfried was supported by scholarships from the Institute of

Marine Biotechnology e.V. and the Ph.D. graduate program of the International Max Planck Research School of Marine Microbiology (MarMic).

Appendix C

Candidatus Prosiliicoccus vernus, a spring phytoplankton bloom associated member of the *Flavobacteriaceae*

T. Ben Francis[1], Karen Krüger[1], Bernhard M. Fuchs[1], Hanno Teeling[1], Rudolf I. Amann[1]

[1]Max Planck Institute for Marine Microbiology, Bremen, Germany

Published in *Systematic and Applied Microbiology* (DOI: 10.1016/j.syapm.2018.08.007).

Contributions to the manuscript:
Experimental concept and design: 15%
Acquisition of experimental data: 10%
Data analysis and interpretation: 15%
Preparation of figures and tables: 0%
Drafting of the manuscript: 0%

Electronic figure versions and supplementary material are available at https://doi.org/10.1016/j.syapm.2018.08.007.

C.1 Abstract

Microbial degradation of algal biomass following spring phytoplankton blooms has been characterised as a concerted effort among multiple clades of heterotrophic bacteria. Despite their significance to overall carbon turnover, many of these clades have resisted cultivation. One clade known from 16S rRNA gene sequencing surveys at Helgoland in the North Sea, was formerly identified as belonging to the genus *Ulvibacter*. This clade rapidly responds to algal blooms, transiently making up as much as 20% of the free-living bacterioplankton. Sequence similarity below 95% between the 16S rRNA genes of described *Ulvibacter* species and those from Helgoland suggest this is a novel genus. Analysis of 40 metagenome assembled genomes (MAGs) derived from samples collected during spring blooms at Helgoland support this conclusion. These MAGs represent three species, only one of which appears to bloom in response to phytoplankton. MAGs with estimated completeness greater than 90% could only be recovered for this abundant species. Additional, less complete, MAGs belonging to all three species were recovered from a mini-metagenome of cells sorted via flow cytometry using the genus specific ULV995 fluorescent rRNA probe. Metabolic reconstruction indicates this highly abundant species most likely degrades proteins and the polysaccharide laminarin. Fluorescence in situ hybridisation showed coccoid cells, with a mean diameter of 0.78 mm, with standard deviation of 0.12 μm. Based on the phylogenetic and genomic characteristics of this clade, we propose the novel candidate genus *Candidatus* Prosiliicoccus, and for the most abundant and well characterised of the three species the name *Candidatus* Prosiliicoccus vernus.

C.2 Introduction

Species of the class *Flavobacteriia* are among the most numerically abundant bacteria in coastal oceans during and immediately following phytoplankton bloom events (Buchan et al., 2014, Chafee et al., 2017, Needham & Fuhrman, 2016, Teeling et al., 2012, 2016). Surveys of bacterial diversity during spring blooms and throughout the year in the North Sea, as well as in other regions of the North Atlantic, have demonstrated repeated patterns of occurrence of specific clades of *Flavobacteriia*, which typically form blooms of their own in response to springtime increases in algal abundance (Chafee et al., 2017, Lindh et al., 2015, Lucas et al., 2015, Teeling et al., 2012, 2016). In common with other Bacteroidetes such as those found in mammalian digestive tracts, it has been proposed that marine *Flavobacteriia* typically make use of higher molecular weight polymeric organic matter such as algal derived protein and polysaccharide (González et al., 2008,

Thomas et al., 2011). Indeed the *Bacteroidetes* are known to frequently possess characteristic genomic structures known as polysaccharide utilisation loci (PULs), which are assemblages of carbohydrate active enzymes (CAZymes) (Terrapon et al., 2017) typically physically co-located in the genome with genes for a TonB-dependent receptor derived transport system (Grondin et al., 2017, Martens et al., 2009, Terrapon et al., 2018, 2015). Many cultivated marine flavobacterial strains such as *Gramella forsetii* KT0803^{T} (Kabisch et al., 2014), *Polaribacter* spp. Hel1_33_49 and Hel1_85 (Xing et al., 2015), and *Zobellia galactanivorans* DsiJ^{T} (Barbeyron et al., 2016b), have been shown to possess multiple PULs in their genomes, which vary in CAZyme content dependent on the specific polysaccharide that is targeted. Detailed description of the degradative capabilities of the free-living spring bloom associated bacterioplankton has, however, largely focused on community-level analysis rather than on individual taxa (Bennke et al., 2016, Teeling et al., 2012, 2016), and currently there are only a handful of cultivated species genuinely representative of the major bloom-forming genera (Hahnke et al., 2015). It is therefore of considerable value to begin describing these bloom associated communities at the level of individual genomes, given that they play a major role in the recycling of organic material and remineralisation of fixed carbon in surface oceans.

One clade identified as an important part of the post-algal bloom flavobacterial community in the North Sea has been previously referred to as belonging to the genus *Ulvibacter* (Chafee et al., 2017, Teeling et al., 2012, 2016). This clade was found to make up as much as 20% of the total free-living bacterial community during and after spring bloom events. Additionally it appeared that it responded concurrent with and in the immediate aftermath of the initial algal bloom, suggesting a possible niche associated with living and senescent algae or their exudates rather than dead algal or bacterial cell material. Recalcitrance of this clade to cultivation has, however, precluded inference of plausible growth substrates, and all that had been known thus far was based on 16S rRNA gene sequences and estimates of environmental abundances measured both by direct cell counting via fluorescence in situ hybridisation (FISH), and by 16S rRNA gene amplicon sequencing.

Here we describe this *Ulvibacter*-related clade as the novel genus *Candidatus* Prosiliicoccus, from the Latin *prosilio* — meaning to jump up, rush, or break forth. The genus currently comprises three North Sea species, one of which is sufficiently well represented in our data that we describe it as a novel species *Candidatus* Prosiliicoccus vernus (henceforth referred to for brevity as *Ca.* Pv), reflecting its initial identification from spring samples. Our circumscription is based on multiple metagenome assembled genomes (MAGs) derived from assembled metagenomic datasets from samples collected across spring blooms in the years 2010, 2011, and 2012. Additional data relating to the

environmental abundance of this genus and its phylogenetic position also support our description.

C.3 Materials and methods

C.3.1 Sampling

Surface seawater samples were collected from the long-term ecological research site Kabeltonne at the island of Helgoland in the North Sea (54° 11.3′ N, 7° 54.0′ E), as described previously (Teeling et al., 2012, 2016). Biomass of free-living bacteria for DNA extraction was collected on 0.2 µm pore-size polycarbonate filters following prefiltration steps using both 10 and 3 µm filters to remove larger primarily eukaryotic material and debris, including particle attached bacteria.

Seawater for FISH was fixed by direct addition of formaldehyde (final concentration of 1%) to the sample, followed by filtration on 0.2 µm pore-size polycarbonate filters without pre-filtration. For cell sorting, 10 l of unfixed water sample from the 21st of April 2009 was filtered onto a polycarbonate filter (142 mm diameter, 0.2 µm pore-size) within 3 h of sampling. All filters were kept frozen at -80 °C until processing.

C.3.2 Fluorescence in situ hybridisation

For visualisation and size estimation, cells of *Ca.* Prosiliicoccus on a filter from 08/04/2010, when it was determined to be highly abundant (Teeling et al., 2016), were labelled using catalysed reporter deposition-FISH (CARD-FISH), using probe ULV995, and following the protocol of (Thiele et al., 2011).

For flow cytometric cell sorting, a modified hybridisation chain reaction-FISH (HCR-FISH) (Yamaguchi et al., 2015) was done, again using ULV995. The initiator probe was the *Ca.* Prosiliicoccus specific ULV-I-995 initiatorH (5'-CCGAATACAAAGCATCAACGACTAGA-AAAA-TCCACGCCTGTCAGACTACA-3'). This was used in conjunction with the two competitors ULV-I-995 c1 (5'-TCCACTCCTGTCAGACTACA-3') and ULV-I-995 c2 (5'-TCCACCCCTGTCAGACTACA-3'). NON EUB initiatorH (5'-CCGAATACAAAGCATCAACGACTAGA-AAAAA-ACTCCTACGGGAGGCAGC-3') was used as a negative control. Hybridisation was done in direct-gene-FISH buffer as described by Barrero-Canosa et al. (2017), and contained 35% formamide with a final probe concentration of 1 µM. The sample was then hybridised for 120 min

at 46 °C, before washing for 30 min at 48 °C in washing buffer. Amplification was carried out with four-times Alexa Fluor™488 – labelled oligonucleotides H1 (5'-TCTAGTCGTT(G)*ATGCTTT(G)*TATTCGGCGACA(G)*ATAACCGAATACAAA(G)*CATC-3') and H2 (5'- CCGAATACAAA(G)*CATCAAC(G)*ACTAGAGATGCTTT(G)*TATTCG(G)*TTATCTGTCG-3') in amplification buffer after they had been denatured for 90 s at 95 °C followed by 30 min at 25 °C, and kept at 20 °C until further use. Amplification was done for 120 min at 37 °C. Final washing was done twice in 1× PBS at 4 °C for 5 min and in deionised water for 30 s. Filters were air dried and embedded in CitiFluor™AF1 supplemented to a final concentration of 2 ng µl-1 with 4',6-diamidino-2-phenylindole (DAPI). Microscopy was done on a Zeiss LSM 780 confocal laser scanning microscope equipped with an Airyscan detector, using a 63× Plan-Apochromat objective lens (Carl Zeiss, Jena, Germany).

C.3.3 Cell sorting using FISH, and sorted cell mini-metagenome generation

A piece of polycarbonate filter containing approximately 8.5 × 10^8 cells (sample date 21/04/2009) was hybridised overnight at 35 °C with the *Ca.* Prosiliicoccus specific ULV-I-995 initiatorH (see *Fluorescence in situ hybridisation* above), washed in washing buffer for 30 min at 35 °C (Yamaguchi et al., 2015) and subsequently incubated in amplification buffer-H1/H2 probe mix in a humidity chamber at 35 °C for 2 h. Hybridised filters were then cut into small pieces, transferred into a 1.5 ml tube containing 1.3 ml of ice-cold cell resuspension buffer (150 mM NaCl, 0.05% Tween80) and vortexed for 15 min at 4 °C. The supernatant containing hybridised cells was kept on ice until analysis and cell sorting. Cell sorting was conducted using a MoFlo flow cytometer (Beckman Coulter, Krefeld, Germany). Supernatants containing resuspended cells were DAPI stained, and prefiltered through a 5 µm pore-size polycarbonate filter (Millipore, 13 mm diameter) to avoid nozzle clogging. Cells were sorted according to their combined FISH and DAPI signal and stored at -20 °C until further processing. The purity of sorted cells was checked microscopically. In order to avoid contamination, DNA amplification from sorted cells was done in a UV-treated PCR workstation using the illustra GenomiPhi V2 DNA Amplification (MDA) Kit (GE Healthcare) according to the manufacturer's instructions. Ten replicates of ∼500 sorted, ULV995-positive cells were lysed by three freeze/thaw cycles (-20 °C/room temperature) and subsequent alkaline lysis before subjection to the MDA reaction. Sorted calibration beads served as a negative control. After taxonomic verification of the MDA products by 16S rRNA gene sequencing, the genome sequencing of the MDA product with the highest yield was performed by JGI using the Illumina MiSeq platform (San Diego, CA, USA) and a 2 × 150 bp protocol. Following sequencing,

raw reads were trimmed and quality filtered to remove the TruSeq adapters and low quality sequence using BBDuk v35.14 (http://bbtools.jgi.doe.gov). Options used for trimming were as follows: ktrim = r k = 28 mink = 12 hdist = 1 tbo = t tpe = t qtrim = rl trimq = 20 minlength = 100. Read quality for each sample was then confirmed using FastQC v0.11.2 (Andrews, 2010).

C.3.4 Metagenome sequencing

Sequencing of ten metagenome samples at the DOE Joint Genome Institute (JGI) has been described previously (Teeling et al., 2016) (sample dates: 03/03/2010; 08/04/2010; 04/05/2010; 18/05/2010; 24/03/2011; 238/04/2011; 26/05/2011; 08/03/2012; 16/04/2012; 10/05/2012). An additional 28 metagenomic samples from intervening dates across the same period were also sequenced at JGI using the same procedures (sample dates: 30/03/2010; 13/04/2010; 20/04/2010; 23/04/2010; 30/04/2010; 11/05/2010; 21/03/2011; 28/03/2011; 31/03/2011; 04/04/2011; 07/04/2011; 14/04/2011; 21/04/2011; 26/04/2011; 06/05/2011; 09/05/2011; 12/05/2011; 16/05/2011; 19/05/2011; 23/05/2011; 30/05/2011; 05/04/2012; 12/04/2012; 26/04/2012; 03/05/2012; 24/05/2012; 31/05/2012; 07/06/2012). For full details of sample preparation and sequencing for each sample see Supplementary Table S1 - online. Briefly, extracted DNA was sheared to average length of 270 bp by sonication, and then sequenced on the Illumina HiSeq platform following a 2 × 150 bp protocol. The ten samples from Teeling et al. (2016) were sequenced more deeply than the 28 additional samples, resulting in approximately four times as many reads. Following sequencing, raw reads were trimmed and quality filtered as detailed above for *Cell sorting using FISH and sorted cell mini-metagenome generation*.

C.3.5 Metagenome assembly and binning

Quality filtered reads from each metagenomic sample and also the sorted cell mini-metagenome were assembled individually using SPAdes v3.10.0 (Nurk et al., 2017) in -meta mode with kmer lengths of 21, 33, 55, 77, and 99, and with read error correction enabled. Contigs longer than 2.5 kbp were retained for binning. Each metagenome assembly was binned using CONCOCT (Alneberg et al., 2014) as part of the standard anvi'o v3 workflow (Eren et al., 2015). To generate differential coverage information for CONCOCT, SPAdes error corrected reads from the assembled sample and the reads from four other randomly selected datasets from the same year were mapped back to the assembly. Reads were mapped with BBMap v35.14 (http://bbtools.jgi.doe.gov), using 'fast' mode, minimum mapping identity (minid) of 0.99, and identity filter for reporting

mappings (idfilter) of 0.97. The sorted cell mini-metagenome was binned directly using the anvi'o interactive interface (anvi-interactive function), using reads mapped in the same manner as above from all 38 metagenomic samples. SPAdes error corrected reads from the sorted mini-metagenome itself were not included as these were the product of a single MDA run and would therefore not be expected to give meaningful differential coverage information between species.

C.3.6 Bin selection and refinement

Bins from metagenomes deriving from *Candidatus* Prosiliicoccus populations were initially identified using the output of CheckM tree v1.0.8 (Parks et al., 2015), which produces an approximate phylogenomic placement of metagenomic bins. Bins were selected for further refinement that had a close phylogenetic relationship to known *Ulvibacter* species and Unidentified eubacterium SCB49, which is also sometimes referred to as *Ulvibacter* sp. SCB49. Additionally, bin similarity was assessed using Mash v1.1.1 (Ondov et al., 2016) with the default sketch size of 1000. A mash distance cutoff of less than 0.05 – approximating average nucleotide identity (ANI) of greater than 95% – was used to determine if two bins belonged to the same species, and this then produced three clusters of bins representing the three *Ca.* Prosiliicoccus species, two of which clusters contained a sorted cell mini-metagenome MAG (metagenome assembled genome). The selected metagenome bins were then manually refined using the anvi'o interactive interface (anvi-refine function) to produce high quality MAGs. In order to produce MAGs with lower L50 and higher N50 values, the refined bins (excluding sorted cell mini-metagenome MAGs) of the *Ca.* Prosiliicoccus species were then "co-reassembled". Reads from each of the 38 metagenomic samples were remapped to each MAG using BBMap as in *Metagenome assembly and binning* above, and all reads mapping to MAGs of the two lower abundant *Ca.* Prosiliicoccus species were then co-assembled with SPAdes in careful mode without error correction enabled using kmers of length 21, 33, 55, 77, 99, and 127. In the case of the more abundant *Ca.* Pv, however, reassembly using reads from all 38 metagenomes did not produce an improved assembly, and so only reads from the 13/04/2010 metagenome dataset that mapped to the most complete MAG (Prosiliicoccus_vernus_Helgoland_20100413) were reassembled, using the same SPAdes parameters as for the other reassemblies. The resulting reassemblies were then refined with anvi'o by mapping reads from each of the 38 metagenomic samples back to the reassembly using BBMap as before, followed by profiling with minimum contig length of 1000 base pairs, and manual refinement in the anvi'o interactive interface. Assessments of MAG completeness and redundancy were made using both CheckM's lineage workflow, with anvi'o, and with the HMM.essential.rb script from the enveomics

collection (Rodriguez-R & Konstantinidis, 2016) in metagenome (-M) mode. The reassembled MAGs were then taken for further analyses.

C.3.7 Phylogenomic and 16S rRNA gene phylogenetic reconstruction

Phylogenomic reconstruction was based on concatenated sequences of 40 proteins (Supplementary Table S2 - online) present in all three reassembled *Ca.* Prosiliicoccus MAGs, taken from the 82 phylogenetically conserved bacterial proteins listed by Soo et al. (2014). Reference amino acid sequences of other *Flavobacteriaceae* and *Bacteroidetes* were downloaded from NCBI GenBank, with the North Sea Gammaproteobacterium *Reinekea* sp. Hel_1_31_D35 used as outgroup. Amino acid sequences were predicted for the *Ca.* Prosiliicoccus MAGs using Prodigal v2.6.3 (Hyatt et al., 2010) in metagenomic mode (-p meta). Sequences of the 40 phylogenetic markers (Supplementary Table S2 - online) were identified in the MAGs and reference genomes using the hmmsearch function of HMMER v3 (Eddy, 2011). Sequences were aligned using FAMSA v1.2 (Deorowicz et al., 2016) with default parameters, and phylogenomic trees calculated using RAxML v8.2.9 (Stamatakis, 2014) with automatic selection of substitution model, and rapid-bootstrapping with 1000 resamplings (-m PROTGAMMAAUTO -p 12345 -x 12345 -# 1000 -o Reinekea_sp_Hel1_31_D35). Trees were visualised using iTOL (Letunic & Bork, 2016).

16S rRNA gene based phylogeny was calculated using the full length 16S rRNA gene sequences detected by anvi'o in the MAGs Prosiliicoccus_vernus_Helgoland_20110523, Prosiliicoccus_vernus_Helgoland_20100518, Prosiliicoccus_vernus_Helgoland_20110421, and Prosiliicoccus_vernus_Helgoland_20110426. These sequences were aligned with SINA v1.3.0 (Pruesse et al., 2012) to the SILVA NR Ref database v128 (Quast et al., 2013), along with all *Flavobacteriaceae* in SILVA v128, and the Bacteroidetes *Bacteroides vulgatus* ATCC 8482 and *Prevotella brevis* ATCC 19188, and using *Reinekea blandensis* MED297^{T} as outgroup. 16S rRNA genes deriving from the two lower abundant *Ca.* Prosiliicoccus species that produced less complete MAGs could not be detected in the assembled metagenomes, and thus could not be included in this part of the analysis. Phylogeny was reconstructed using RAxML with the same bootstrapping as above, but using the GTRGAMMA substitution model (-f a -m GTRGAMMA -p12345 -x 12345 -# 1000 -o Reinekea_blandensis_MED297).

Phylogenetic uniqueness was assessed using both percent identity across the full length 16S rRNA gene sequences, calculated in ARB (Ludwig et al., 2004), and by calculation of ANI and average amino acid identity (AAI) between MAGs and the genomes of related

species using the ani.rb and aai.rb scripts from the enveomics collection (Rodriguez-R & Konstantinidis, 2016).

C.3.8 Estimation of environmental abundance

Direct cell counting to estimate absolute cell numbers using CARD-FISH has been described previously (Teeling et al., 2016). Cell counts from that study made using hybridisation with the ULV995 probe, which we consider to be specific to the genus *Ca.* Prosiliicoccus, were used here to plot absolute cell numbers. The relevant data from that work is reproduced here in Supplementary Table S3 - online.

Similarly, estimates of relative abundance based on proportion of reads deriving from amplicon data from samples also collected at Helgoland have been described previously (Chafee et al., 2017). Data from Chafee et al. (2017) referring to *Ulvibacter* are used here for additional monitoring of this clade. Sequence identity between the most abundant oligotype sequence classified by Chafee et al. (2017) as *Ulvibacter* and the 16S rRNA gene from the *Ca.* Prosiliicoccus MAGs was 100%.

Data for global abundance and distribution was collected using IMNGS (Lagkouvardos et al., 2016), using the 16S rRNA gene sequence from the MAG Prosiliicoccus_vernus_Helgoland_20110426 as a query. Minimum target size was 200, and an identity threshold of 99% was used. Percent of reads in each sequencing run was calculated from the IMNGS output, and the corresponding geographic positions for each sequencing run (where these data were available) were collected from NCBI. An arbitrary cutoff of at at least 50 reads matching the query was used for plotting.

Species relative abundance was also assessed based on the proportion of metagenomic reads recruited to individual bins. Reads from all 38 metagenomic samples were thus mapped to the reassembled MAGs of each *Ca.* Prosiliicoccus species, and the number of reads recruited were counted and normalised to the total number of reads in that sample. These numbers then estimate the proportion of reads deriving from the different *Ca.* Prosiliicoccus populations over time. Reads were mapped as detailed above in *Metagenome assembly and binning.*

C.3.9 Assessment of single nucleotide variation and strain diversity in *Ca.* Prosiliicoccus vernus

Single nucleotide variants (SNVs) in all MAGs, including the reassembled MAGs, were called by anvi'o using the anvi-gen-variability-profile tool, with -min-coverage-in-each-sample set to 20×. Metagenomic samples were selected for inclusion in this analysis

based on the average detection of the MAG by each sample, as calculated by anvi'o. Samples where detection of the MAG was greater than or equal to 0.9 were included, meaning that at least 90% of the nucleotide positions in the MAG had at least one read mapping to them. Number of SNVs per thousand base pairs and average coverage of SNVs in each MAG were then calculated from the output.

Inference of number and abundance of strains represented by the reassembled *Ca.* Pv MAG was also attempted using DESMAN (Quince et al., 2017) as described on the DESMAN github pages (https://github.com/chrisquince/DESMAN). The mapping files created for measuring abundance were used as inputs. Core COGs used were the 40 identified by Mende et al. (2013). The variant filter was run with -p and -c options, and the coverage cutoff was reduced to 2 in order to include more samples from early 2011 and 2012. The same coverage cutoff was used for running the DESMAN algorithm. The DESMAN algorithm was run with 10 replicates, with -r 1000 and -i 500 as recommended.

C.3.10 MAG annotation and metabolic reconstruction

Initial annotation was done with Prokka v1.12 (Seemann, 2014), modified to include prediction and annotation of partial genes by removing the -c and -m options when running Prodigal within Prokka. This annotation was then manually refined for *Ca.* Pv using searches against Pfam v31.0 (Finn et al., 2016) (pfam_scan.pl script with default parameters), and BLAST v2.6.0+ (Altschul et al., 1997) searches against the most up-to-date NCBI nr database (downloaded 13/03/2018). For all three species, specific annotation of CAZymes using the dbCAN v6 database (Yin et al., 2012) and the hmmscan function of HMMER was also used. Custom e-value cutoffs for specific CAZyme families were used as described previously (Teeling et al., 2016). Comparison of CAZyme sequence identity against experimentally verified sequences was done using BLAST. Peptidases were predicted using BLAST against the MEROPS merops_scan database v12.0 (Rawlings et al., 2012), using the default BLAST settings recommended by MEROPS: e-value cutoff of 1×10^{-4}. Cellular localisation of glycoside hydrolase and peptidase enzymes was predicted using CELLO v2.5 (Yu et al., 2006) and PSORTb v3.0 (Yu et al., 2010). SusC/D-like transporters were predicted using the TIGRFAM (Haft et al., 2013) profile TIGR04056 for SusC-like sequences and Pfam profiles PF07980.9, PF12741.5, PF12771.5, and PF14322.4 for SusD-like sequences. Additional TonB dependent receptors were predicted using TIGRFAM profiles TIGR01352, TIGR01776, TIGR01778, TIGR01779, TIGR01782, TIGR01783, TIGR01785, TIGR01786, TIGR02796, TIGR02797, TIGR02803, TIGR02804, TIGR02805, and TIGR04057. For TonB- dependent receptors and SusC-like genes, an e-value cutoff of 1×10^{-10} was used, and for SusD-like genes an e-value cutoff of 1×10^{-5} was used. Function of sulfatases was

predicted using SulfAtlas v1.2 (Barbeyron et al., 2016a). Metabolic pathway information was reconstructed from the Prokka output using Pathway Tools v20.5 (Karp et al., 2016).

C.3.11 Data availability

Metagenome reads for the 38 environmental metagenomes used in this study are available under the NCBI BioProject accession numbers listed in Supplementary Table S1 - online. The sorted cell mini-metagenome reads are available under NCBI BioProject accession PRJNA367155. Accession numbers for the metagenome assemblies and *Ca.* Prosiliicoccus MAG sequences were deposited in ENA using the data brokerage service of the German Federation for Biological Data (GFBio) (Diepenbroek et al., 2014) in compliance with the Minimal Information about any (X) Sequence (MIxS) standard (Yilmaz et al., 2011), and are available under the INSDC project number PRJEB28156. Anvi'o databases for individual MAGs, and the sorted cell mini-metagenome assembly, sorted MAGs, and sorted cell metagenome anvi'o database are available at doi:10.6084/m9.figshare.6139730.

C.4 Results

C.4.1 Metagenomic sequencing

Summary information for the metagenomic datasets from the dates 03/03/2010; 08/04/2010; 04/05/2010; 18/05/2010; 24/03/2011; 28/04/2011; 26/05/2011; 08/03/2012; 16/04/2012; and 10/05/2012 have been reported previously (Teeling et al., 2016), and are reproduced in the data in Supplementary Table S4 - online along with summary statistics covering the rest of the metagenomic samples.

C.4.2 Metagenome assembly and binning

General assembly statistics for each metagenomic dataset are presented in Supplementary Table S4 - online. The binning and MAG reassembly process produced the 40 environmental metagenome derived MAGs and seven sorted cell mini-metagenome derived MAGs described in Supplementary Table S5 - online. The MAGs divided into three clusters we consider to represent three distinct species (Fig. 3.1, Fig. 3.2a), as determined by both phylogenomic placement and average nucleotide identity. Redundancy of single copy marker genes of 3% or below for all MAGs indicates low levels of

contamination. Estimates of completeness and redundancy as calculated by CheckM, anvi'o, and the HMM.essential.rb script of the enveomics collection are also included in Supplementary Table S5 - online. Approximately 75% of mini-metagenome reads mapped back to the mini-metagenome derived *Ca.* Prosiliicoccus MAG sequences at 97% identity, with a ratio of approximately 770:120:1 between *Ca.* Pv and the second and third species respectively.

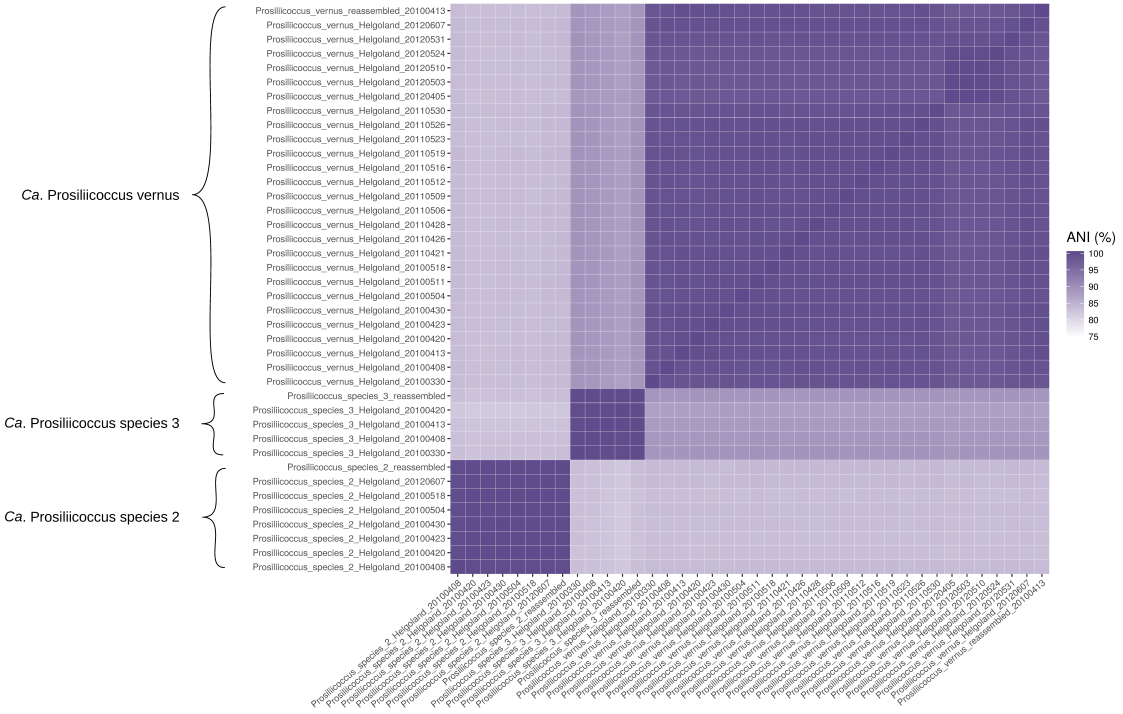


Figure 3.1: Average nucleotide identity between *Candidatus* Prosiliicoccus MAGs showing separation into three species.

The reassembled MAGs showed improvements in total length and completeness without any increase in redundancy (see Supplementary Table S5 - online). The reassembled MAG of *Ca.* Pv has completeness estimated at between 92% and 96%, with redundancy between 0.4% and 0.7% and total length of 1.9 Mbp. The reassembled MAG of *Ca.* Prosiliicoccus species 2 (*Ca.* P2) is between 69% and 91% complete with 0–0.6% contamination and length of 1.9 Mbp, while the reassembled MAG of *Ca.* Prosiliicoccus species 3 (*Ca.* P3) is between 42% and 64% complete with 0% redundancy and length of 1.35 Mbp. Note that CheckM gave the reassembled MAG of *Ca.* P2 a completeness score of above 90%, which may be considered a threshold for completeness upon which description of a candidate species may be based, while the other two approaches to measure completeness did not. Contamination/redundancy estimates were effectively nil in both CheckM (0.63%) and anvi'o (0%). Since the consensus between metrics was not

in favour of the MAG being near complete however, the MAG is not formally described here as a candidate species.

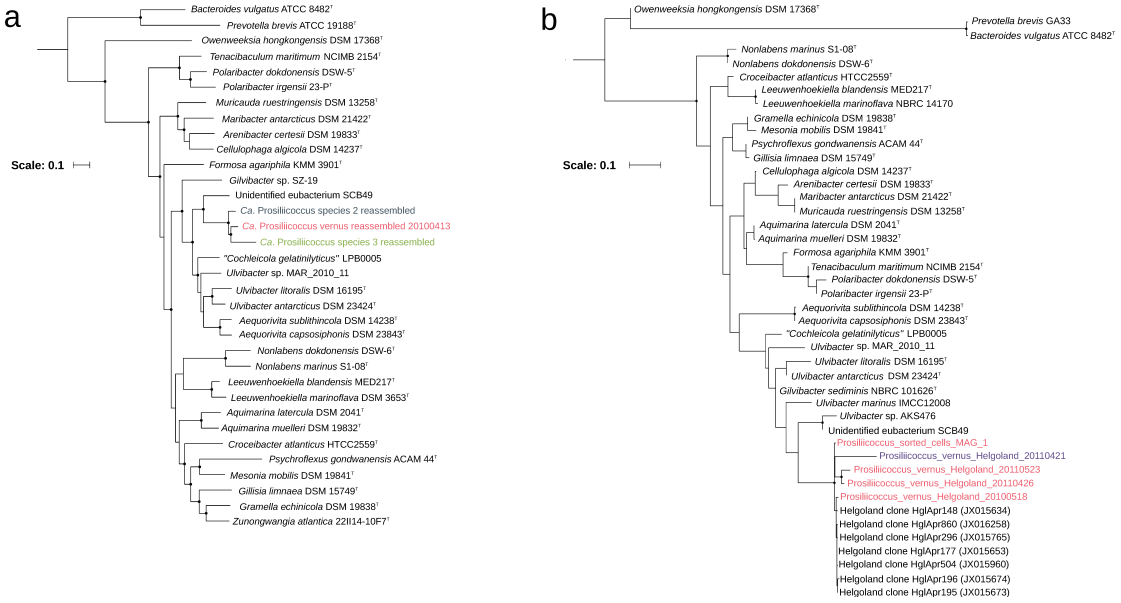


Figure 3.2: Phylogeny of *Ca.* Prosiliicoccus. (**a**) Phylogenomic reconstruction based on 40 concatenated proteins (Supplementary Table S2 - online). (**b**) 16S rRNA gene based phylogenetic reconstruction. Both phylogenies were calculated using RAxML with 1000 rapid bootstrap replicates. Black dots on nodes indicate greater than 90% bootstrap support.

Average nucleotide identities between MAGs and reassembled MAGs for the three *Ca.* Prosiliicoccus species are shown in Fig. 3.1 and Supplementary Table S6 - online, indicating clear genomic divergence between the three groups. ANI between *Ca.* P2 and the other two species was approximately 84%, while ANI between *Ca.* P3 and *Ca.* Pv was higher at approximately 90%.

While MAGs 1 and 2 from the sorted cell mini-metagenome clearly belong to *Ca.* Pv and *Ca.* P2 respectively, as was predicted by Mash, sorted MAG 7 has ANI of greater than 98% to *Ca.* P3, while 3, 6, 9, and 10 all appear to belong within the genus (Supplementary Fig. S1 - online). The clear caveat with these MAGs is that they each had lengths below 500 kbp and were not complete enough to be clustered by Mash. The low completeness of these MAGs is expected to be the result of the MDA not amplifying all parts of genomes equally. Despite this, the sorted cell mini-metagenome derived MAGs confirm the connection between 16S rRNA gene sequences, the ULV995 probe, and our metagenome derived MAGs.

Genomic divergence between the *Ca.* Pv MAGs was higher than that detected between the two lower abundant species, implying greater strain diversity in the *Ca.* Pv population. This can be seen in the higher proportion of genomic fragments with lower

ANI (Supplementary Fig. S2 - online) and the lower ANI values between whole MAGs (minimum value 97.7% between the *Ca.* Pv MAGs from 30/03/2010 and 05/04/2012, as compared to minima above 99% within species 2 and species 3 MAG sets), as can be seen in Supplementary Fig. S1 - online and Supplementary Table S6 - online. It is also apparent that there is a change in the *Ca.* Pv population between non-bloom and bloom time periods, as demonstrated by the lower ANI between MAGs from earlier in 2012 and those assembled from the final two samples from 2012, which coincided with the onset of the *Ca.* Pv bloom (ANI between first and final *Ca.* Pv MAG from 2012 = 98.0%, compared to ANI of above 99% between the first four MAGs from 2012, and similarly between the final pair). The pre-bloom population from 2012 is also seemingly more genetically homogeneous than the bloom populations from 2010, 2011 and 2012, implying a change in population composition during progression of the 2012 bloom (i.e. the first four MAGs from 2012 have ANI above 99.7% with one another, compared with lower values between MAGs from consecutive sampling dates from the previous years).

ANI and AAI between the reassembled *Ca.* Prosiliicoccus MAGs and reference genomes of related species are shown in Supplementary Fig. S3 - online and Supplementary Table S7 - online, indicating that *Ca.* Prosiliicoccus species belong to a separate genus with higher ANI and AAI within the genus than between *Ca.* Prosiliicoccus and other genera. ANI between MAGs within the genus *Ca.* Prosiliicoccus is approximately 84% and above, compared to values below 80% when compared to genomes of other species. AAI values within the genus are approximately 88% and above, compared to below 70% when compared with reference taxa.

Four of the *Ca.* Pv MAGs contained near full length 16S rRNA genes (Prosiliicoccus_vernus_Helgoland_20110426, Prosiliicoccus_vernus_Helgoland_20110523, Prosiliicoccus_vernus_Helgoland_20100518, and Prosiliicoccus_vernus_Helgoland_20110421), as did Prosiliicoccus_sorted_MAG_1. A further three (Prosiliicoccus_vernus_Helgoland_20100413, Prosiliicoccus_vernus_Helgoland_20110509, and Prosiliicoccus_vernus_Helgoland_20110530) contained other parts of the rRNA operon. From visual inspection of the coverage of contigs containing these operons, there appears to be approximately twofold higher coverage of the rRNA operon than the rest of the contig on which they sit, implying the presence of most likely two rRNA operons in this organism.

C.4.3 Phylogenomic and 16S rRNA phylogenetic reconstruction

Analysis of 40 concatenated, phylogenetically conserved protein sequences indicates the sister group relationship between *Ca.* Prosiliicoccus and *Ulvibacter*, as well as the presence of three distinct species within the genus *Ca.* Prosiliicoccus (Fig. 3.2a). This is consistent with the 16S rRNA gene based phylogeny, which recovers the same relationship between *Ulvibacter* and *Ca.* Prosiliicoccus (Fig. 3.2b). Equally, both methods are clear on the placement of *Ca.* Prosiliicoccus in the family *Flavobacteriaceae*, order *Flavobacteriales*, and class *Flavobacteriia* in the phylum *Bacteroidetes*. As is evident from the ANI data, the phylogenomic reconstruction confirms that *Ca.* Pv and *Ca.* P3 are more closely related to one another than either is to *Ca.* P2.

16S rRNA gene identity, typically used to determine the taxonomic level of divergence between clades, also confirms that *Ca.* Prosiliicoccus belongs to a novel genus, as identity between the full length *Ca.* Pv 16S rRNA sequence (using that derived from MAG Prosiliicoccus_vernus_Helgoland_20100518 as representative) and those of *Gilvibacter sediminis* NBRC 101626 (Khan et al., 2007) (94.1%), *Ulvibacter antarcticus* DSM 23424^{T} (Choi et al., 2007) (93.7%), and *Ulvibacter litoralis* DSM 16195^{T} (Nedashkovskaya et al., 2004) (94.7%) lie close to and below the threshold for delineating genera as recommended by Yarza et al. (2014). The similarity between *Ca.* Pv and Unidentified eubacterium SCB49 (sometimes referred to as *Ulvibacter* sp. SCB49) is 94.9%, thus also at the lower bound of belonging to genus *Ca.* Prosiliicoccus, with both the phylogenomic and 16S rRNA based reconstructions placing it as the closest relative of the three *Ca.* Prosiliicoccus species presented here.

The 16S rRNA gene assembled in the MAG Prosiliicoccus_vernus_Helgoland_20110421 only shares identity of 97% with those from the other *Ca.* Pv MAGs, and it is most likely that this sequence derives from misassembly of the gene because the coverage profile of the rest of the contig on which this gene is found is consistent with *Ca.* Pv. It appears from the 16S rRNA phylogeny that the genus *Ulvibacter* may also be paraphyletic, with *Ulvibacter marinus* IMCC 12008^{T} (Baek et al., 2014) belonging to a sister group to the other two described *Ulvibacter* species. Similarly the isolate *Ulvibacter* sp. MAR_2010_11 is likely a member of another genus, as is demonstrated in both phylogenies.

C.4.4 Cell morphology

In epifluorescence microscopic images of *Ca.* Prosiliicoccus, identified using CARD-FISH with the ULV995 16S rRNA probe, cells appear coccoid (Supplementary Fig. S4 - online).

In contrast, all species of *Ulvibacter* so far described are rods (Baek et al., 2014, Choi et al., 2007, Nedashkovskaya et al., 2004). *Ca.* Prosiliicoccus cells have a size range of approximately 0.5-1 µm in diameter (Supplementary Table S8 - online) with the mean of all cells measured being 0.78 µm, and a standard deviation of 0.12 µm.

C.4.5 Estimates of environmental abundance

In the years 2010, 2011, and 2012, *Ca.* Prosiliicoccus species reached high abundances during and after phytoplankton blooms (Fig. 3.3). Rapid doubling times are possible for the *Ca.* Prosiliicoccus population; a minimum doubling time of less than one day, implied by greater than 100% daily increases in cell number, can be seen at certain time points in Fig. 3.3a and Supplementary Table S3 - online. This data refers to the population of the entire genus however, given that the ULV995 probe targets all three species.

From the amplicon data from Chafee et al. (2017), it is apparent that at least in 2010 and 2011, *Ca.* Prosiliicoccus populations were prevalent during both spring and summer phytoplankton blooms at Helgoland, suggesting that conditions that favour this clade are not restricted to the springtime (Fig. 3.3b). This is also backed up by the global pattern of detection (Fig. 3.4), which demonstrates that sequences with high identity to the *Ca.* Pv 16S rRNA gene have been detected in regions such as the Benguela upwelling system off the coast of Namibia, and at a site in the Southern Ocean where an artificial iron seeded phytoplankton bloom was generated (Thiele et al., 2012), as well as the seasonal temperate northern hemisphere locations where it can be seen in high abundance (Supplementary Table S9 - online). Additionally, sequences have been detected in lower abundance across a number of sites in both the northern and southern hemispheres, demonstrating the ubiquity of this clade in temperate and polar regions. It is likely these data refer to the genus *Ca.* Prosiliicoccus as a whole, as based on our inability to distinguish the three species in the amplicon data of Chafee et al. (2017), we might expect standard 16S rRNA gene amplicon datasets to capture a region of this gene conserved across the three species.

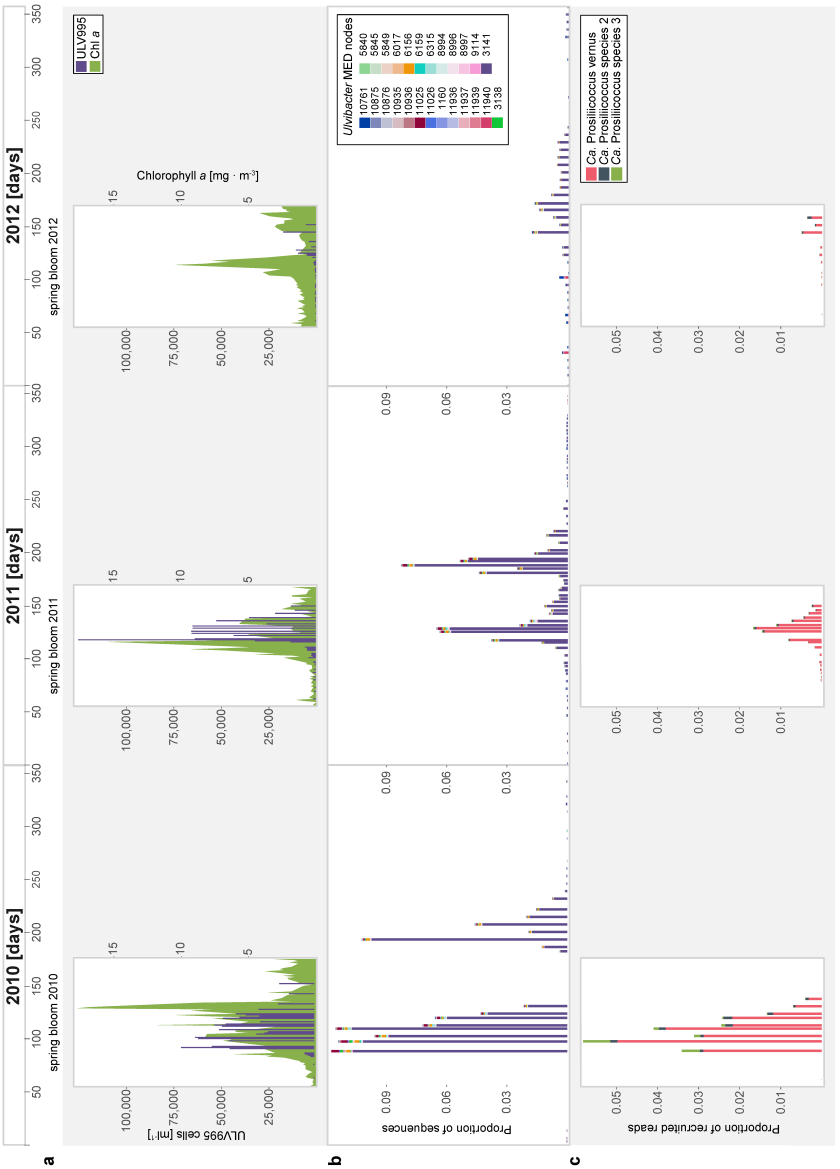


Figure 3.3: Estimates of environmental abundance of *Ca.* Prosiliicoccus in the years 2010-12 at Helgoland during spring blooms, based on (**a**) CARD-FISH cell counts and Chlorophyll a data from Teeling et al. (2016); (**b**) proportion of amplicon sequences classified as *Uvibacter* in the data from Chafee et al. (2017); (**c**) proportion of reads from each of the 38 metagenomic datasets recruited to the reassembled *Ca.* Prosiliicoccus MAGs.

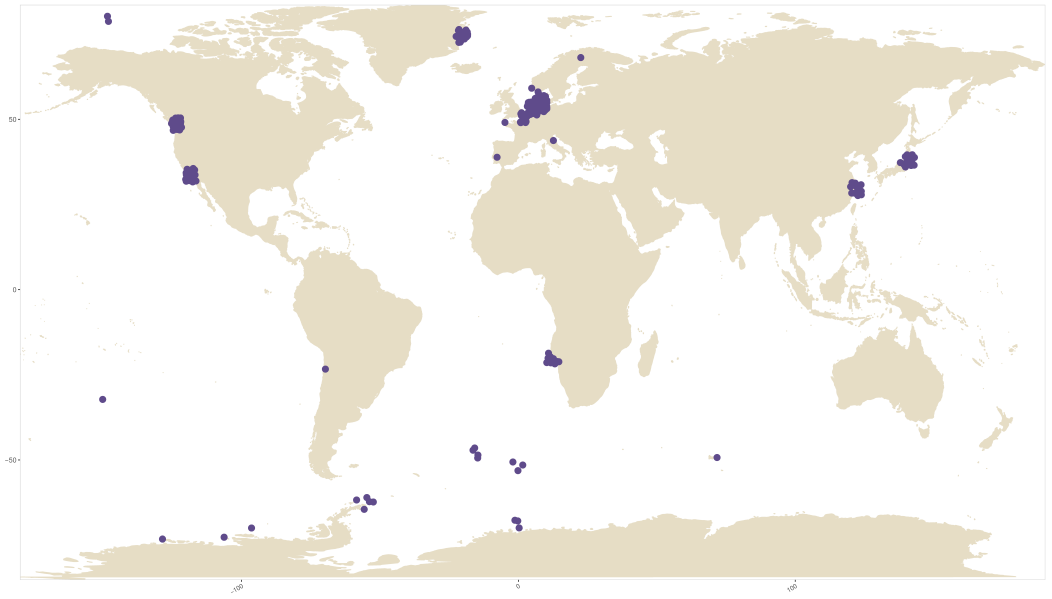


Figure 3.4: Global distribution of amplicons showing greater than 99% identity to the 16S rRNA gene of *Ca.* Prosiliicoccus vernus. Points indicate detection, with an arbitrary cutoff of at least 50 reads from the sample matching the query sequence. The sample in the Pacific Ocean derives from a sponge metagenome (*Scopalina* sp.; BioProject accession PRJNA292036). Points are jittered to demonstrate where multiple sampling of the same location has taken place.

The metagenomic read recruitment data from Helgoland shows a similar pattern of abundance of the three *Ca.* Prosiliicoccus species to the other two datasets (Fig. 3.3c). It is apparent that *Ca.* Pv is the dominant *Ca.* Prosiliicoccus species during spring blooms, and based on the differential in average sequencing depth between *Ca.* Pv and the two lower abundant *Ca.* Prosiliicoccus species, it most likely makes up the majority of the *Ca.* Prosiliicoccus community detected by FISH and amplicon sequencing.

C.4.6 Within species variation in *Ca.* Prosiliicoccus populations

While sequencing depth was not sufficient for meaningful information to be gleaned regarding the two lower abundant *Ca.* Prosiliicoccus species in 2011 and 2012, variation within all three species in 2010 could be examined using SNV profiling. The detection of different strains within populations represented by individual MAGs was also attempted. As was seen when comparing ANI between MAGs, variability was higher within the *Ca.* Pv population than within the populations of the other two species (Supplementary Fig. S5 - online), with the number of variable nucleotide positions per kilobase pair typically between 5 and 10, as compared to less than 3 in species 2 and species 3. The exception is the pre-bloom phase of 2012, where, as was seen in the ANI data, detected variability was lower in the *Ca.* Pv population. Strain deconvolution produced only inconclusive results. DESMAN's built in heuristic for strain number prediction suggested

that *Ca.* Pv comprises 5 confidently predicted strains. However, visual inspection of the mean posterior deviance (Supplementary Fig. S6 - online), which should stop decreasing once the true number of strains has been modelled, suggests that even at 14 strains the curve was only beginning to look as though it might plateau. This suggests the variability in this population is either genuinely very high, or it is in some way inconsistent with the underlying principles of the DESMAN algorithm. When using the run predicted as optimal by DESMAN (9 haplotypes used, 5 of which were confidently predicted, with average error of 5%) to examine strain relative abundance in the *Ca.* Pv population however, it appears that the strains that are predicted have consistent abundance patterns across the bloom periods, implying a deterministic separation of function between strain-like populations of *Ca.* Pv despite the high noise (Supplementary Fig. S7 - online). This consistency between years can also be seen when using data from runs with higher predicted strain numbers (data not shown). What is evident from Supplementary Fig. S7 - online is that haplotype H2 is dominant in 2010 and 2011 (∼50% of the total *Ca.* Pv population), and as the *Ca.* Pv population increases in size through 2012, this haplotype increases to approach a similar proportion of the overall population. This suggests that haplotype H2 could be responding more strongly to the phytoplankton blooms than other haplotypes. The dominance of haplotype H3 before bloom onset in 2012 is also consistent with the homogeneity seen among MAGs from these dates based on ANI. There is also a noticeable shift in both 2010 and 2011 from H5 to H4 over time, again pointing potentially to deterministic rather than stochastic changes in population structure. These patterns are currently only observed in these 3 years however, and this conclusion is thus only tentative.

C.4.7 Annotation of the reassembled *Ca.* Prosiliicoccus vernus MAG and inference of metabolic potential

The reassembled *Ca.* Pv MAG contains 1810 predicted genes, 592 of which (33%) are annotated as hypothetical. Of these, 31 are tRNA genes, among which tRNA genes for aspartic acid are absent. tRNA genes for aspartic acid are however found in other *Ca.* Pv MAGs.

C.4.8 Basic energy conservation

The *Ca.* Pv MAG contains complete pathways for aerobic respiration comprising glycolysis, the non-oxidative phase of the pentose phosphate pathway, TCA cycle, and an electron transport chain (Fig. 3.5). There are no unambiguous indications of use of other monosaccharides than glucose, but the presence of various unspecified ABC transporters

implies that different sugar monomers might also be taken up. The MAG also possesses 87 predicted proteases, some of which are expected to be secreted extracellularly (Supplementary Table S10 - online), and predicted degradation pathways are present for the amino acids alanine, arginine, asparagine, cysteine, glutamine, histidine, isoleucine, lysine, methionine, phenylalanine, threonine, tryptophan, tyrosine, and valine. Additionally the MAG contains 10 co-located genes involved in phenylacetate degradation, which encode the multisubunit 1,2-phenylacetyl-CoA epoxidase (PaaABCDE) as well as PaaGHINY enzymes. There is also a gene for a short chain fatty acid transporter, long chain fatty acid ligase, and an alkane 1-monooxygenase gene, indicating basic processing of fatty acids either as an additional source of reduced carbon or for general metabolic purposes such as building of cell membranes.

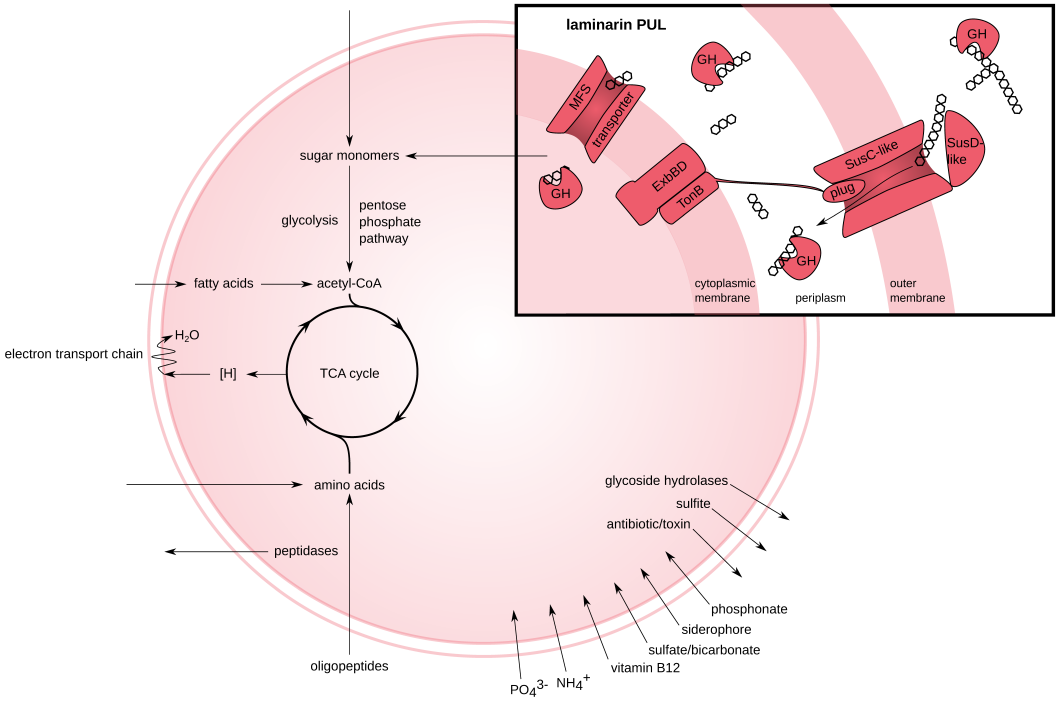


Figure 3.5: General summary of predicted metabolic potential of *Ca.* Prosiliicoccus vernus. Energy conservation occurs via consumption of low molecular weight organic molecules such as small peptides and glycans, and the abundant polysaccharide laminarin. Other sugar monomers than glucose could not be definitively shown to be utilised. Light energy can also be conserved via the use of a proteorhodopsin. Otherwise *Ca.* Prosiliicoccus vernus is an obligate aerobe which uses Na^+ translocating NADH-quinone reductase as part of its respiratory chain.

The MAG contains genes for all 6 subunits of the Na^+-translocating NADH-quinone oxidoreductase, succinate dehydrogenase, and cytochrome c and cytochrome c oxidases of complex IV. In our data, there are no indications of adaptation for fermentation, or for use of alternative electron acceptors for anaerobic respiration.

The presence of a glycogen synthase gene suggests energy storage. Also present is a proteorhodopsin gene, which could be involved either in energy conservation, or otherwise may serve as part of a light based sensory system.

C.4.9 Sources of nitrogen, sulfur, phosphorous

The primary nitrogen source for *Ca.* Pv is expected to be protein and amino acids, whilst it also possesses an ammonium transporter. There is no indication of assimilatory sulfate reduction in the genome, so it is presumed that proteins are the primary sulfur source for this species. There are transporters present for both inorganic phosphate and organic phosphonates, and both polyphosphate kinase and exopolyphosphatase genes that imply storage of phosphate via polyphosphate.

C.4.10 Transport

Ion transporters are well represented in the reassembled *Ca.* Pv MAG, with transporters for copper, magnesium, zinc, sodium, potassium, cobalt, manganese, iron (including probable siderophore carrying TonB-dependent transporters), and bicarbonate/sulfate, in addition to the aforementioned ammonium transporter.

Also present are transporters belonging to the *gld* family found in many flavobacteria, and also to the type IX secretion system. Other genes with predicted export function include homologues of multidrug exporters and macrolide exporters, and a fluoride efflux transporter. Import functions include di/tripeptide transporters, an oligopeptide permease, the aforementioned fatty acid transporter, various TonB dependent transporters including SusC-like homologues, and ABC and MFS type transporters. These transporters can have diverse functions that are not readily predicted with regard to their substrate but are expected to include transporters of cobalamin, which *Ca.* Pv appears to be auxotrophic for. One interesting SusC/D-like pair in this genome is not co-located with any CAZymes, and thus does not have an inferred function in polysaccharide transport. Instead it sits in close proximity to a collection of ribosomal protein genes, a transcription elongation factor, and an outer membrane protein assembly factor. The function of this pair remains indeterminate.

C.4.11 CAZyme profile and predicted polysaccharide utilisation

Annotation of CAZymes and transporters indicated the presence of a single canonical polysaccharide utilisation structure. This PUL includes a SusC/D-like pair, and otherwise contains just a pair of glycoside hydrolases of the GH16 and GH3 families (Fig. 3.6),

in close proximity to genes for ATP-dependent 6-phosphofructokinase, glyceraldehyde-phosphate dehydrogenase, and glucosephosphate isomerase (all involved in glycolysis). The GH16 enzyme is predicted to be extracellular, while the GH3 is predicted to be either periplasmic or extracellular (Supplementary Table S10 - online). The most probable putative substrate for this PUL is laminarin, as most characterised GH3 family enzymes have β-glucosidase activity, and GH16 family members, while having more widely varying described functions, are known to include β-1,3-glucanase activity that would act on laminarin. Additional support for this comes from GH3- and GH16-containing PULs that have been experimentally confirmed to be laminarin-specific (Kabisch et al., 2014). Identity between protein sequences of the GH16 in *Ca.* Prosiliicoccus vernus and its homologue in the experimentally confirmed PUL from *Gramella forsetii* KT0803^{T} (Kabisch et al., 2014) was 40%, and identity between GH3s was 58%. The fact that the proteins belong to the same families, and share similar neighbouring gene functions such as the SusC/D-like pair, support the prediction of laminarin as the substrate of this PUL, despite the fact that specific functional domains could not be assigned to the sequences. On the same contig, some 30 kbp removed from the SusC/D-like pair, the reassembled MAG contains a pair of GH17 enzymes, also known to be active on laminarin and to be part of laminarin active PULs (Becker et al., 2017). The *Ca.* Pv MAG Prosiliicoccus_vernus_Helgoland_20100420 has a more complete assembly of this region of the genome, and possesses a cluster including three GH17 family genes (two cytoplasmic, the other undetermined), together with one GH30 and one GH2 gene (no consensus on localisation), and an MFS family glucan transporter that is localised to the cytoplasmic membrane (Supplementary Table S10 - online). This is identical in gene content and order to that found in the *Ca.* P2 reassembled MAG contig 37 (Fig. 3.6). *Ca.* Pv also has a gene pair comprising a GH20 *beta*-hexosaminidase and a GH2 exo-*beta*-glucosaminidase. Glucosamine is typically found in chitin and chitosan, two other abundant marine polysaccharides. There is no gene for any N-acetylglucosamine transporter in close proximity to these genes in any of the *Ca.* Pv MAGs however.

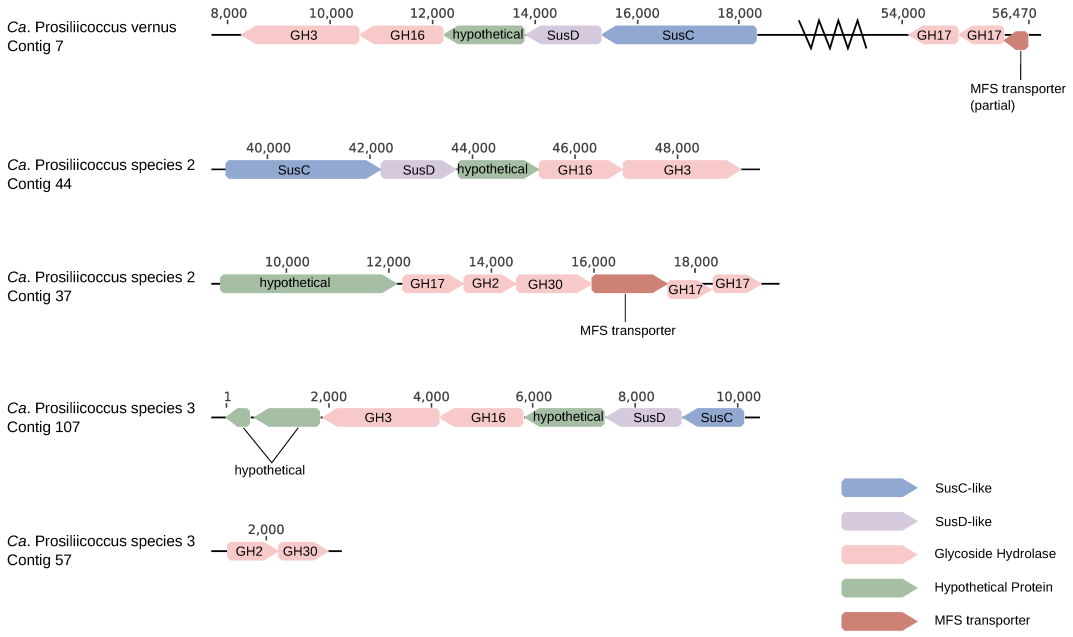


Figure 3.6: Putative laminarin degrading PUL-like structures in reassembled *Ca.* Prosiliicoccus MAGs, showing conserved gene order among the three species. The two types of structure seen here, the SusC/D-like pair with GH16 and GH3, and the GH17/GH30 collections, are both similar to laminarin degrading PULs known in other North Sea species. Numbers above genes indicate position on contig in base pairs.

C.4.12 Annotation of reassembled *Ca.* P2 and P3 MAGs and partial inference of metabolic potential

The reassembled *Ca.* P2 MAG contains 1765 predicted genes, with 690 of those predicted to be hypothetical, and 32 tRNAs, of which only tRNA genes for isoleucine and tryptophan are absent. The predicted metabolic capabilities are similar to that for *Ca.* Pv, with only the absence of cysteine and lysine degradation, and fatty acid metabolism being noteworthy. With respect to CAZyme profile, *Ca.* P2 has the same 3 CAZyme collections mentioned for *Ca.* Pv, namely the GH3, GH16, SusC/D-like pair and GH17, GH30 collections putatively involved in laminarin degradation shown in Fig. 3.6, and the GH2, GH20 pair plausibly involved in chitin utilisation.

The reassembled MAG of *Ca.* P3 contains 1306 predicted genes, including 536 hypothetical genes and 21 tRNAs. As with *Ca.* P2, the predicted metabolic pathways are similar to those for *Ca.* Pv, although with fewer complete pathways and more absent predicted pathways, as is to be expected from a less complete genome. The CAZyme and PUL profile of *Ca.* P3 differs from that of the other two species. While it has the same SusC/D-like pair, GH16, GH3 PUL and at least a hint of the second laminarin degrading cluster contained on a short contig with a GH30 and GH2 gene (Fig. 3.6), it also possesses a gene cluster containing a GH29 family enzyme and two sulfatases

(one family S1_19 endo-carageenan function, the other family S1_25, unknown function), a PUL with SusC/D-like pair and two GH86 family β-porphyranases/-agarases putatively degrading porphyran or agar, and a PUL that consists of a SusC/D-like pair, three predicted sulfatase genes (families S1_36 and S1_16 - unknown function, and family S1_11 – heparan/mucin function), a GH128 family gene, and two hypothetical proteins (Supplementary Fig. S8 - online). The implication from these gene collections, despite the fact that they may not be completely assembled here, is that there exists a distinct substrate niche for this species that includes sulfated polysaccharides which are not available to the two other *Ca.* Prosiliicoccus species.

C.5 Discussion

The novel candidate genus *Candidatus* Prosiliicoccus, family *Flavobacteriaceae*, order *Flavobacteriales*, class *Flavobacteriia*, phylum *Bacteroidetes* presented here comprises three distinct species detected in metagenomic datasets deriving from samples collected during spring blooms in the North Sea at the island of Helgoland. The genus *Ca.* Prosiliicoccus comprises apparently obligate aerobic heterotrophs, which react to phytoplankton blooms. One species is more abundant during blooms than the other two, and near complete metagenome assembled genomes could be recovered that describe this population. Based on these and associated data presented here, we formally describe the candidate species *Candidatus* Prosiliicoccus vernus, according to the standards outlined by Konstantinidis et al. (Konstantinidis & Rossello-Mora, 2015, Konstantinidis et al., 2017).

Candidatus Prosiliicoccus vernus is capable of rapid growth to the extent that, based on FISH counts, population doubling times can approach and at times even fall below one day. Because of this growth, this species can react swiftly to phytoplankton blooms and transiently make up between 5-20% of the total free-living bacterioplankton population. Rapid growth may be in part facilitated by the smaller genome and concomitant small cell size when compared to other known *Flavobacteriia*. *Ca.* Pv most likely relies on a combination of protein, small peptide, and free amino acids as a primary source of carbon and nitrogen. Additionally it is likely that it consumes some form of the polysaccharide laminarin, a storage polysaccharide produced by diatoms and brown algae. Laminarin is released in large amounts during spring blooms at Helgoland, as a result of the high abundance of diatoms during these periods (Teeling et al., 2016). The use of laminarin by *Ca.* Prosiliicoccus vernus is therefore likely to be substantial. Global abundance patterns indicate this species is restricted to temperate and polar latitudes, but appears to respond to phytoplankton bloom events in many locations where they are known to

occur. This implies a commonality between phytoplankton blooms that can be exploited by *Ca.* Prosiliicoccus populations, although the nature of this commonality is not yet determined.

What is unfortunately still unclear from our data, is what precisely permits *Ca.* Prosiliicoccus vernus populations, and indeed perhaps only three strains of this species, to respond so strongly to the increases in algal abundance. This is a particularly challenging question to answer, given that the recovered gene content of the three *Ca.* Prosiliicoccus species is generally similar (corroborated by the high amino acid identity between the species), and that the temporal abundance pattern is similar for all three species despite the vast disparities in cell numbers. It is unsurprising that three species evidently capable of consuming laminarin and protein would respond to some extent to the massive increases in these substrates in the water column that occur as a result of phytoplankton blooms, thus some other mechanism not here readily determinable is necessary to explain the size of *Ca.* Pv populations during the spring at Helgoland. The existence of homologues of nutrient uptake systems (specifically the phosphate and ammonia transporters, and the polyphosphate kinase and exopolyphosphatase) in *Ca.* P2 suggests that presence of these gene functions alone is also unlikely to be allowing the high growth rates of *Ca.* Pv populations.

Compared to the closely related genera *Ulvibacter*, *Aequorivita*, *Altibacter* and "*Cochleicola*", and the genome of Unidentified eubacterium SCB49, *Candidatus* Prosiliicoccus vernus has several notable differences. Firstly the genome is much reduced in size, at 1.9 Mbp compared to the typical 3-4 Mbp genomes of close relatives. Cells are also coccoid rather than rod shaped, which sets them apart not only from close relatives, but also from other free living *Flavobacteriia* known to be abundant in the North Sea such as members of the genera *Formosa* and *Polaribacter*. It has been suggested that reduced cell size and coccoid morphology can have adaptive benefits in evading grazers (Pernthaler, 2005), which might aid the rapid growth capacity of *Ca.* Pv. The putatively reduced capacity for degradation of complex polysaccharides, when compared to many *Flavobacteriia*, including others known to respond to phytoplankton blooms in the North Sea, is an additional distinguishing feature of *Ca.* Pv.

The multiple lines of evidence presented here, including divergence of 16S rRNA gene sequences, divergence of conserved single copy genes, genomic distinctness in terms of ANI and AAI, and the recovery and description of a near complete metagenome assembled genome, all support the description of a novel species and genus within the *Flavobacteriaceae*.

C.5.1 Description of *Candidatus* Prosiliicoccus

Candidatus Prosiliicoccus (Pro.si.li.i.coc'cus. L. v. *prosilio*, *prosilire*, *prosilui* to leap, jump, rush, spring forth; N.L. mas. n. *coccus* from Gr. mas. n. *kokkos* grain, seed; N.L. mas. n. *Prosiliicoccus*).

Members of the genus *Ca.* Prosiliicoccus are predicted to be obligate aerobes, with currently no indication of fermentation or anaerobic respiration. They are heterotrophic, marine surface water dwelling bacteria, capable of using glycans and proteins as primary sources of organic matter. The three species have been detected in seawater sampled in the North Sea during spring phytoplankton blooms via metagenomic assembly and fluorescence in situ hybridisation. Cells are coccoid, and may be detected using FISH probe ULV995 (Teeling et al., 2012). High quality 16S rRNA gene sequences share approximately 94% identity with closely related genera *Ulvibacter* and *Gilvibacter*. G + C content for all three species is between 36% and 37%. The genus *Ca.* Prosiliicoccus belongs to the family *Flavobacteriaceae*, order *Flavobacteriales*, class *Flavobacteriia*, and phylum *Bacteroidetes*. Type species is *Candidatus* Prosiliicoccus vernus.

C.5.2 Description of *Candidatus* Prosiliicoccus vernus

Ca. Prosiliicoccus vernus (ver'nus. L. mas. adj. *vernus* pertaining to spring, vernal).

Genome annotation allows prediction of consumption of protein, peptides, and amino acids, as well as putatively the polysaccharide laminarin. The spectrum of glycans putatively available to *Candidatus* Prosiliicoccus vernus is restricted, with laminarin appearing to be the most significant polysaccharide, while it may also consume chitin. Genome size is small relative to described *Flavobacteriia*, at an estimated 1.9 Mbp. Fluorescence microscopy reveals the cells to be coccoid, with diameter ranging between 0.5 and 1 μm. Bloom-forming behaviour is observed in *Candidatus* Prosiliicoccus vernus during and immediately after phytoplankton blooms in the North Sea.

Type material is the metagenome assembled genome 'Prosiliicoccus_vernus_reassembled_20100413' submitted to ENA in project PRJEB28156, and also to the Digital Protologue database under TaxoNumber CA00022. Together the data presented here fulfil all of the criteria required for description of uncultivated prokaryotic taxa outlined by Konstantinidis et al. (2017).

C.6 Acknowledgements

The authors acknowledge the support of the Max Planck Society. TBF, BMF, HT, and RIA acknowledge funding from German Research Foundation (DFG) project FOR 2406 – 'Proteogenomics of Marine Polysaccharide Utilisation (POMPU)'. Metagenome sequencing was conducted in the framework of the COGITO project (Contract No. DE-AC02-05CH11231) by the U.S. Department of Energy Joint Genome Institute, a DOE Office of Science User Facility, and is supported by the Office of Science of the U.S. Department of Energy under Contract No. DE-AC02-05CH11231. The authors acknowledge the contribution of Tijana Glavina del Rio as project leader of metagenome sequencing at JGI on the COGITO project. We also thank Aharon Oren for his expertise and input in matters of nomenclature, and Ivaylo Kostadinov of GFBio for support with deposition of sequence data.

Bibliography

Adams, P. D., Afonine, P. V., Bunkóczi, G., Chen, V. B., Davis, I. W. et al. (2010). PHENIX: a comprehensive Python-based system for macromolecular structure solution. *Acta Crystallographica Section D Biological Crystallography*, **66**, 213–221.

Alderkamp, A.-C., Buma, A. G. J. & van Rijssel, M. (2007a). The carbohydrates of *Phaeocystis* and their degradation in the microbial food web. *Biogeochemistry*, **83**, 99–118.

Alderkamp, A.-C., van Rijssel, M. & Bolhuis, H. (2007b). Characterization of marine bacteria and the activity of their enzyme systems involved in degradation of the algal storage glucan laminarin. *FEMS Microbiology Ecology*, **59**, 108–117.

Ale, M. T., Mikkelsen, J. D. & Meyer, A. S. (2011). Important determinants for fucoidan bioactivity: a critical review of structure-function relations and extraction methods for fucose-containing sulfated polysaccharides from brown seaweeds. *Marine Drugs*, **9**, 2106–2130.

Allègre, C., Manhès, G. & Lewin, E. (2001). Chemical composition of the Earth and the volatility control on planetary genetics. *Earth and Planetary Science Letters*, **185**, 49–69.

Alneberg, J., Bjarnason, B. S., de Bruijn, I., Schirmer, M., Quick, J., Ijaz, U. Z., Lahti, L., Loman, N. J., Andersson, A. F. & Quince, C. (2014). Binning metagenomic contigs by coverage and composition. *Nature Methods*, **11**, 1144–1146.

Alonso, C., Warnecke, F., Amann, R. & Pernthaler, J. (2007). High local and global diversity of *Flavobacteria* in marine plankton. *Environmental Microbiology*, **9**, 1253–1266.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. (1990). Basic Local Alignment Search Tool. *Journal of Molecular Biology*, **215**, 403–410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J. H., Zhang, Z., Miller, W. & Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new

generation of protein database search programs. *Nucleic Acids Research*, **25**, 3389–3402.

Amin, S. A., Hmelo, L. R., van Tol, H. M., Durham, B. P., Carlson, L. T., Heal, K. R., Morales, R. L., Berthiaume, C. T., Parker, M. S., Djunaedi, B., Ingalls, A. E., Parsek, M. R., Moran, M. A. & Armbrust, E. V. (2015). Interaction and signalling between a cosmopolitan phytoplankton and associated bacteria. *Nature*, **522**, 98–101.

Amin, S. A., Parker, M. S. & Armbrust, E. V. (2012). Interactions between diatoms and bacteria. *Microbiology and Molecular Biology Reviews*, **76**, 667–684.

Anderson, K. L. & Salyers, A. A. (1989). Genetic evidence that outer membrane binding of starch is required for starch utilization by *Bacteroides thetaiotaomicron*. *Journal of Bacteriology*, **171**, 3199–3204.

Andersson, A. F., Riemann, L. & Bertilsson, S. (2010). Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *The ISME Journal*, **4**, 171–181.

Andrews, S. (2010). FastQC: a quality control tool for high throughput sequence data. http://www.bioinformatics.babraham.ac.uk/projects/fastqc.

Apprill, A., McNally, S., Parsons, R. & Weber, L. (2015). Minor revision to V4 region SSU rRNA 806R gene primer greatly increases detection of SAR11 bacterioplankton. *Aquatic Microbial Ecology*, **75**, 129–137.

Arlov, Ø., Aachmann, F. L., Sundan, A., Espevik, T. & Skjåk-Bræk, G. (2014). Heparin-like properties of sulfated alginates with defined sequences and sulfation degrees. *Biomacromolecules*, **15**, 2744–2750.

Arlov, Ø. & Skjåk-Bræk, G. (2017). Sulfated alginates as heparin analogues: a review of chemical and functional properties. *Molecules*, **22**, 778.

Armbrust, E. V. (2009). The life of diatoms in the world's oceans. *Nature*, **459**, 185–192.

Armstrong, Z., Mewis, K., Liu, F., Morgan-Lang, C., Scofield, M., Durno, E., Chen, H. M., Mehr, K., Withers, S. G. & Hallam, S. J. (2018). Metagenomics reveals functional synergy and novel polysaccharide utilization loci in the *Castor canadensis* fecal microbiome. *The ISME Journal*, **12**, 2757–2769.

Arnosti, C. (2011). Microbial extracellular enzymes and the marine carbon cycle. *Annual Review of Marine Science*, **3**, 401–425.

Arnosti, C., Durkin, S. & Jeffrey, W. H. (2005). Patterns of extracellular enzyme activities among pelagic marine microbial communities: implications for cycling of dissolved organic carbon. *Aquatic Microbial Ecology*, **38**, 135–145.

Arnosti, C., Fuchs, B. M., Amann, R. & Passow, U. (2012). Contrasting extracellular enzyme activities of particle-associated bacteria from distinct provinces of the North Atlantic Ocean. *Frontiers in Microbiology*, **3**, 425.

Arrieta, J. M., Mayol, E., Hansman, R. L., Herndl, G. J., Dittmar, T. & Duarte, C. M. (2015). Dilution limits dissolved organic carbon utilization in the deep ocean. *Science*, **348**, 331–333.

Aspeborg, H., Coutinho, P. M., Wang, Y., Brumer, Harry, I. & Henrissat, B. (2012). Evolution, substrate specificity and subfamily classification of glycoside hydrolase family 5 (GH5). *BMC Evolutionary Biology*, **12**, 186.

Avci, B., Hahnke, R. L., Chafee, M., Fischer, T., Gruber-Vodicka, H., Tegetmeyer, H. E., Harder, J., Fuchs, B. M., Amann, R. I. & Teeling, H. (2017). Genomic and physiological analyses of '*Reinekea forsetii*' reveal a versatile opportunistic lifestyle during spring algae blooms. *Environmental Microbiology*, **19**, 1209–1221.

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

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T. et al. (2008). The RAST Server: rapid annotations using subsystems technology. *BMC Genomics*, **9**, 75.

Baek, K., Jo, H., Choi, A., Kang, I. & Cho, J. C. (2014). *Ulvibacter marinus* sp nov., isolated from coastal seawater. *International Journal of Systematic and Evolutionary Microbiology*, **64**, 2041–2046.

Barbeyron, T., Brillet-Guéguen, L., Carré, W., Carrière, C., Caron, C., Czjzek, M., Hoebeke, M. & Michel, G. (2016a). Matching the diversity of sulfated biomolecules: creation of a classification database for sulfatases reflecting their substrate specificity. *PLoS ONE*, **11**, e0164846.

Barbeyron, T., Thomas, F., Barbe, V., Teeling, H., Schenowitz, C., Dossat, C., Goesmann, A., Leblanc, C., Oliver Glöckner, F., Czjzek, M., Amann, R. & Michel, G. (2016b). Habitat and taxon as driving forces of carbohydrate catabolism in marine heterotrophic bacteria: example of the model algae-associated bacterium *Zobellia galactanivorans* Dsij^{T}. *Environmental Microbiology*, **18**, 4610–4627.

Barras, D. R. & Stone, B. A. (1969). Beta-1,3-glucan hydrolases from *Euglena gracilis*. II. Purification and properties of the beta-1,3-glucan exo-hydrolase. *Biochimica Et Biophysica Acta*, **191**, 342–353.

Barrero-Canosa, J., Moraru, C., Zeugner, L., Fuchs, B. M. & Amann, R. (2017). Direct-geneFISH: a simplified protocol for the simultaneous detection and quantification of genes and rRNA in microorganisms. *Environmental Microbiology*, **19**, 70–82.

Bauer, M., Kube, M., Teeling, H., Richter, M., Lombardot, T. et al. (2006). Whole genome analysis of the marine *Bacteroidetes* '*Gramella forsetii*' reveals adaptations to degradation of polymeric organic matter. *Environmental Microbiology*, **8**, 2201–2213.

Beattie, A., Percival, E. & Hirst, E. L. (1961). Studies on metabolism of *Chrysophyceae*. Comparative structural investigations on leucosin (chrysolaminarin) separated from diatoms and laminarin from brown algae. *Biochemical Journal*, **79**, 531–537.

Becker, S., Scheffel, A., Polz, M. F. & Hehemann, J.-H. (2017). Accurate quantification of laminarin in marine organic matter with enzymes from marine microbes. *Applied and Environmental Microbiology*, **83**, e03389–16.

Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., Jovanovich, S. B., Gates, C. M., Feldman, R. A., Spudich, J. L., Spudich, E. N. & DeLong, E. F. (2000a). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, **289**, 1902–1906.

Béjà, O., Suzuki, M. T., Koonin, E. V., Aravind, L., Hadd, A., Nguyen, L. P., Villacorta, R., Amjadi, M., Garrigues, C., Jovanovich, S. B., Feldman, R. A. & DeLong, E. F. (2000b). Construction and analysis of bacterial artificial chromosome libraries from a marine microbial assemblage. *Environmental Microbiology*, **2**, 516–529.

Bell, R. T. & Kuparinen, J. (1984). Assessing phytoplankton and bacterioplankton production during early spring in Lake Erken, Sweden. *Applied and Environmental Microbiology*, **48**, 1221–1230.

Benner, R., Louchouarn, P. & Amon, R. M. W. (2005). Terrigenous dissolved organic matter in the Arctic Ocean and its transport to surface and deep waters of the North Atlantic. *Global Biogeochemical Cycles*, **19**, https://doi.org/10.1029/2004GB002398.

Bennke, C. M., Krüger, K., Kappelmann, L., Huang, S., Gobet, A., Schüler, M., Barbe, V., Fuchs, B. M., Michel, G., Teeling, H. & Amann, R. I. (2016). Polysaccharide utilisation loci of *Bacteroidetes* from two contrasting open ocean sites in the North Atlantic. *Environmental Microbiology*, **18**, 4456–4470.

Bennke, C. M., Neu, T. R., Fuchs, B. M. & Amann, R. (2013). Mapping glycoconjugate-mediated interactions of marine *Bacteroidetes* with diatoms. *Systematic and Applied Microbiology*, **36**, 417–425.

Bernardet, J.-F., Segers, P., Vancanneyt, M., Berthe, F., Kersters, K. & Vandamme, P. (1996). Cutting a Gordian knot: emended classification and description of the genus *Flavobacterium*, emended description of the family *Flavobacteriaceae*, and proposal of *Flavobacterium hydatis* nom. nov. (basonym, *Cytophaga aquatilis* Strohl and Tait 1978). *International Journal of Systematic Bacteriology*, **46**, 128–148.

Bersch, M. (1995). On the circulation of the northeastern North-Atlantic. *Deep-Sea Research I*, **42**, 1583–1607.

Bjursell, M. K., Martens, E. C. & Gordon, J. I. (2006). Functional genomic and metabolic studies of the adaptations of a prominent adult human gut symbiont, *Bacteroides thetaiotaomicron*, to the suckling period. *Journal of Biological Chemistry*, **281**, 36269–36279.

Bowers, R. M., Kyrpides, N. C., Stepanauskas, R., Harmon-Smith, M., Doud, D. et al. (2017). Minimum information about a single amplified genome (MISAG) and a metagenome-assembled genome (MIMAG) of bacteria and archaea. *Nature Biotechnology*, **35**, 725–731.

Bowman, J. P. (2000). Description of *Cellulophaga algicola* sp nov., isolated from the surfaces of Antarctic algae, and reclassification of *Cytophaga uliginosa* (ZoBell and Upham 1944) Reichenbach 1989 as *Cellulophaga uliginosa* comb. nov. *International Journal of Systematic and Evolutionary Microbiology*, **50**, 1861–1868.

Bowman, J. P., McCammon, S. A., Lewis, T., Skerratt, J. H., Brown, J. L., Nichols, D. S. & McMeekin, T. A. (1998). *Psychroflexus torquis* gen. nov., sp. nov., a psychrophilic species from Antarctic sea ice, and reclassification of *Flavobacterium gondwanense* (Dobson et al. 1993) as *Psychroflexus gondwanense* gen. nov., comb. nov. *Microbiology*, **144 (Pt 6)**, 1601–1609.

Brown, C. T., Hug, L. A., Thomas, B. C., Sharon, I., Castelle, C. J., Singh, A., Wilkins, M. J., Wrighton, K. C., Williams, K. H. & Banfield, J. F. (2015). Unusual biology across a group comprising more than 15% of domain Bacteria. *Nature*, **523**, 208–211.

Brum, J. R., Morris, J. J., Décima, M. & Stukel, M. R. (2014). Mortality in the oceans: causes and consequences. In P. F. Kemp, ed., *Eco-DAS IX Symposium Proceedings*, 16–48. ASLO.

Bryson, S., Li, Z., Chavez, F., Weber, P. K., Pett-Ridge, J., Hettich, R. L., Pan, C., Mayali, X. & Mueller, R. S. (2017). Phylogenetically conserved resource partitioning in the coastal microbial loop. *The ISME Journal*, **11**, 2781–2792.

Buchan, A., LeCleir, G. R., Gulvik, C. A. & González, J. M. (2014). Master recyclers: features and functions of bacteria associated with phytoplankton blooms. *Nature Reviews Microbiology*, **12**, 686–698.

Buchfink, B., Xie, C. & Huson, D. H. (2014). Fast and sensitive protein alignment using DIAMOND. *Nature Methods*, **12**, 59–60.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V. & Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Research*, **37**, D233–D238.

Caporaso, J. G., Lauber, C. L., Walters, W. A., Berg-Lyons, D., Lozupone, C. A., Turnbaugh, P. J., Fierer, N. & Knight, R. (2011). Global patterns of 16S rRNA diversity at a depth of millions of sequences per sample. *Proceedings of the National Academy of Sciences of the United States of America*, **108**, 4516–4522.

Cardman, Z., Arnosti, C., Durbin, A., Ziervogel, K., Cox, C., Steen, A. D. & Teske, A. (2014). *Verrucomicrobia* are candidates for polysaccharide-degrading bacterioplankton in an Arctic Fjord of Svalbard. *Applied and Environmental Microbiology*, **80**, 3749–3756.

Chafee, M., Fernàndez-Guerra, A., Buttigieg, P. L., Gerdts, G., Eren, A. M., Teeling, H. & Amann, R. I. (2017). Recurrent patterns of microdiversity in a temperate coastal marine environment. *The ISME Journal*, **12**, 237–252.

Chiovitti, A., Harper, R. E., Willis, A., Bacic, A., Mulvaney, P. & Wetherbee, R. (2005). Variations in the substituted 3-linked mannans closely associated with the silicified walls of diatoms. *Journal of Phycology*, **41**, 1154–1161.

Chiovitti, A., Higgins, M. J., Harper, R. E., Wetherbee, R. & Bacic, A. (2003). The complex polysaccharides of the raphid diatom *Pinnularia viridis* (*Bacillariophyceae*). *Journal of Phycology*, **39**, 543–554.

Cho, K. H. & Salyers, A. A. (2001). Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by *Bacteroides thetaiotaomicron*. *Journal of Bacteriology*, **183**, 7224–7230.

Choi, T. H., Lee, H. K., Lee, K. & Cho, J. C. (2007). *Ulvibacter antarcticus* sp nov., isolated from Antarctic coastal seawater. *International Journal of Systematic and Evolutionary Microbiology*, **57**, 2922–2925.

Chow, C.-E. T., Sachdeva, R., Cram, J. A., Steele, J. A., Needham, D. M., Patel, A., Parada, A. E. & Fuhrman, J. A. (2013). Temporal variability and coherence of euphotic zone bacterial communities over a decade in the Southern California Bight. *The ISME Journal*, **7**, 2259–2273.

Cockburn, D. W. & Koropatkin, N. M. (2016). Polysaccharide degradation by the intestinal microbiota and its influence on human health and disease. *Journal of Molecular Biology*, **428**, 3230–3252.

Corrigan, A. J. & Robyt, J. F. (1979). Nature of the fructan of *Streptococcus*-mutans OMZ 176. *Infection and Immunity*, **26**, 387–389.

Cottrell, M. T. & Kirchman, D. L. (2000). Natural assemblages of marine proteobacteria and members of the *Cytophaga-Flavobacter* cluster consuming low- and high-molecular-weight dissolved organic matter. *Applied and Environmental Microbiology*, **66**, 1692–1697.

Cottrell, M. T., Yu, L. Y. & Kirchman, D. L. (2005). Sequence and expression analyses of *Cytophaga*-like hydrolases in a Western arctic metagenomic library and the Sargasso Sea. *Applied and Environmental Microbiology*, **71**, 8506–8513.

Cowtan, K. (2006). The Buccaneer software for automated model building. 1. Tracing protein chains. *Acta Crystallographica Section D Biological Crystallography*, **62**, 1002–1011.

Cox, J. & Mann, M. (2008). MaxQuant enables high peptide identification rates, individualized p.p.b.-range mass accuracies and proteome-wide protein quantification. *Nature Biotechnology*, **26**, 1367–1372.

Coyne, M. J., Zitomersky, N. L., McGuire, A. M., Earl, A. M. & Comstock, L. E. (2014). Evidence of extensive DNA transfer between *Bacteroidales* species within the human gut. *mBio*, **5**, e01305–14.

Cram, J. A., Chow, C.-E. T., Sachdeva, R., Needham, D. M., Parada, A. E., Steele, J. A. & Fuhrman, J. A. (2014). Seasonal and interannual variability of the marine bacterioplankton community throughout the water column over ten years. *The ISME Journal*, **9**, 563–580.

Cuskin, F., Lowe, E. C., Temple, M. J., Zhu, Y., Cameron, E. A. et al. (2015). Human gut *Bacteroidetes* can utilize yeast mannan through a selfish mechanism. *Nature*, **517**, 165–169.

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

de Jesus Raposo, M. F., de Morais, A. M. M. B. & de Morais, R. M. S. C. (2014). Bioactivity and applications of polysaccharides from marine microalgae. In K. Ramawat & J. Mérillon, eds., *Polysaccharides*, 1–38. Springer.

Delmont, T. O., Eren, A. M., Vineis, J. H. & Post, A. F. (2015). Genome reconstructions indicate the partitioning of ecological functions inside a phytoplankton bloom in the Amundsen Sea, Antarctica. *Frontiers in Microbiology*, **6**, 1090.

Delmont, T. O., Quince, C., Shaiber, A., Esen, O. C., Lee, S. T. M., Rappe, M. S., McLellan, S. L., Lücker, S. & Eren, A. M. (2018). Nitrogen-fixing populations of *Planctomycetes* and *Proteobacteria* are abundant in surface ocean metagenomes. *Nature Microbiology*, **3**, 804–813.

DeLong, E. F., Franks, D. G. & Alldredge, A. L. (1993). Phylogenetic diversity of aggregate-attached vs free-living marine bacterial assemblages. *Limnology and Oceanography*, **38**, 924–934.

Deniaud, E., Fleurence, J. & Lahaye, M. (2003). Interactions of the mix-linked beta-(1,3)/beta-(1,4)-D-xylans in the cell walls of *Palmaria palmata* (*Rhodophyta*). *Journal of Phycology*, **39**, 74–82.

Deorowicz, S., Debudaj-Grabysz, A. & Gudyś, A. (2016). FAMSA: Fast and accurate multiple sequence alignment of huge protein families. *Scientific Reports*, **6**, 33964.

Diepenbroek, M., Glöckner, F. O., Grobe, P., Güntsch, A., Huber, R., König-Ries, B., Kostadinov, I., Nieschulze, J., Seeger, B., Tolksdorf, R. & Triebel, D. (2014). Towards an integrated biodiversity and ecological research data management and archiving platform: the German federation for the curation of biological data (GFBio). In E. Plödereder, L. Grunske, E. Schneider & D. Ull, eds., *Informatik 2014*, 1711–1721. Gesellschaft für Informatik e.V.

Dittmar, T. & Stubbins, A. (2014). Dissolved Organic Matter in Aquatic Systems. In H. D. Holland & K. K. Turekian, eds., *Treatise on Geochemistry, Vol 12: Organic Geochemistry*, 125–156. Elsevier.

Eddy, S. R. (2011). Accelerated Profile HMM Searches. *PLoS Computational Biology*, **7**, e1002195.

Edgar, R. C. (2010). Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460–2461.

Eklöf, J. M., Shojania, S., Okon, M., McIntosh, L. P. & Brumer, H. (2013). Structure-function analysis of a broad specificity *Populus trichocarpa* endo-beta-glucanase reveals an evolutionary link between bacterial licheninases and plant XTH gene products. *Journal of Biological Chemistry*, **288**, 15786–15799.

El Kaoutari, A., Armougom, F., Gordon, J. I., Raoult, D. & Henrissat, B. (2013). The abundance and variety of carbohydrate-active enzymes in the human gut microbiota. *Nature Reviews Microbiology*, **11**, 497–504.

Eren, A. M., Esen, O. C., Quince, C., Vineis, J. H., Morrison, H. G., Sogin, M. L. & Delmont, T. O. (2015). Anvi'o: an advanced analysis and visualization platform for 'omics data. *PeerJ*, **3**, e1319.

Falkowski, P. G., Barber, R. T. & Smetacek, V. (1998). Biogeochemical controls and feedbacks on ocean primary production. *Science*, **281**, 200–206.

Fernández-Gómez, B., Richter, M., Schüler, M., Pinhassi, J., Acinas, S. G., González, J. M. & Pedrós-Alió, C. (2013). Ecology of marine *Bacteroidetes*: a comparative genomics approach. *The ISME Journal*, **7**, 1026–1037.

Ficko-Blean, E., Préchoux, A., Thomas, F., Rochat, T., Larocque, R. et al. (2017). Carrageenan catabolism is encoded by a complex regulon in marine heterotrophic bacteria. *Nature Communications*, **8**, 1685.

Field, C. B., Behrenfeld, M. J., Randerson, J. T. & Falkowski, P. (1998). Primary production of the biosphere: integrating terrestrial and oceanic components. *Science*, **281**, 237–240.

Finn, R. D., Bateman, A., Clements, J., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Heger, A., Hetherington, K., Holm, L., Mistry, J., Sonnhammer, E. L. L., Tate, J. & Punta, M. (2014). Pfam: the protein families database. *Nucleic Acids Research*, **42**, D222–D230.

Finn, R. D., Coggill, P., Eberhardt, R. Y., Eddy, S. R., Mistry, J., Mitchell, A. L., Potter, S. C., Punta, M., Qureshi, M., Sangrador-Vegas, A., Salazar, G. A., Tate, J. & Bateman, A. (2016). The Pfam protein families database: towards a more sustainable future. *Nucleic Acids Research*, **44**, D279–D285.

Finn, R. D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J. E., Gavin, O. L., Gunasekaran, P., Ceric, G., Forslund, K., Holm, L.,

Sonnhammer, E. L. L., Eddy, S. R. & Bateman, A. (2010). The Pfam protein families database. *Nucleic Acids Research*, **38**, D211–D222.

Fischer, F. G. & Dörfel, H. (1955). Die Polyuronsäuren der Braunalgen (Kohlenhydrate der Algen I). *Hoppe-Seyler´s Zeitschrift für physiologische Chemie*, **302**, 186–203.

Foley, M. H., Martens, E. C. & Koropatkin, N. M. (2018). SusE facilitates starch uptake independent of starch binding in *B. thetaiotaomicron*. *Molecular Microbiology*, **108**, 551–566.

Ford, C. & Percival, E. (1965). 1299. Carbohydrates of *Phaeodactylum tricornutum*. Part II. A sulphated glucuronomannan. *Journal of the Chemical Society*, **0**, 7042–7046.

Francis, T. B., Krüger, K., Fuchs, B. M., Teeling, H. & Amann, R. I. (2018). *Candidatus* Prosiliicoccus vernus, a spring phytoplankton bloom associated member of the *Flavobacteriaceae*. *Systematic and Applied Microbiology*, https://doi.org/10.1016/j.syapm.2018.08.007.

Fuhrman, J. A., Cram, J. A. & Needham, D. M. (2015). Marine microbial community dynamics and their ecological interpretation. *Nature Reviews Microbiology*, **13**, 133–146.

Garcia-Vaquero, M., Rajauria, G., O'Doherty, J. V. & Sweeney, T. (2017). Polysaccharides from macroalgae: recent advances, innovative technologies and challenges in extraction and purification. *Food Research International*, **99**, 1011–1020.

Gerlach, W. & Stoye, J. (2011). Taxonomic classification of metagenomic shotgun sequences with CARMA3. *Nucleic Acids Research*, **39**, e91.

Ghiglione, J. F., Conan, P. & Pujo-Pay, M. (2009). Diversity of total and active free-living vs. particle-attached bacteria in the euphotic zone of the NW Mediterranean Sea. *FEMS Microbiology Letters*, **299**, 9–21.

Giebel, H.-A., Kalhoefer, D., Lemke, A., Thole, S., Gahl-Janssen, R., Simon, M. & Brinkhoff, T. (2011). Distribution of Roseobacter RCA and SAR11 lineages in the North Sea and characteristics of an abundant RCA isolate. *The ISME Journal*, **5**, 8–19.

Gilbert, J. A., Steele, J. A., Caporaso, J. G., Steinbrueck, L., Reeder, J., Temperton, B., Huse, S., McHardy, A. C., Knight, R., Joint, I., Somerfield, P., Fuhrman, J. A. & Field, D. (2012). Defining seasonal marine microbial community dynamics. *The ISME Journal*, **6**, 298–308.

Giovannoni, S. J., Tripp, H. J., Givan, S., Podar, M., Vergin, K. L., Baptista, D., Bibbs, L., Eads, J., Richardson, T. H., Noordewier, M., Rappe, M. S., Short, J. M., Carrington, J. C. & Mathur, E. J. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, **309**, 1242–1245.

Glenwright, A. J., Pothula, K. R., Bhamidimarri, S. P., Chorev, D. S., Baslé, A., Firbank, S. J., Zheng, H., Robinson, C. V., Winterhalter, M., Kleinekathöfer, U., Bolam, D. N. & van den Berg, B. (2017). Structural basis for nutrient acquisition by dominant members of the human gut microbiota. *Nature*, **541**, 407–411.

Gobet, A., Barbeyron, T., Matard-Mann, M., Magdelenat, G., Vallenet, D., Duchaud, E. & Michel, G. (2018). Evolutionary evidence of algal polysaccharide degradation acquisition by *Pseudoalteromonas carrageenovora* 9^{T} to adapt to macroalgal niches. *Frontiers in Microbiology*, **9**, 2740.

González, J. M., Fernandez-Gomez, B., Fernandez-Guerra, A., Gomez-Consarnau, L., Sanchez, O., Coll-Llado, M., del Campo, J., Escudero, L., Rodriguez-Martinez, R., Alonso-Saez, L., Latasa, M., Paulsen, I., Nedashkovskaya, O., Lekunberri, I., Pinhassi, J. & Pedros-Alio, C. (2008). Genome analysis of the proteorhodopsin-containing marine bacterium *Polaribacter* sp. MED152 (*Flavobacteria*). *Proceedings of the National Academy of Sciences of the United States of America*, **105**, 8724–8729.

Goris, J., Konstantinidis, K. T., Klappenbach, J. A., Coenye, T., Vandamme, P. & Tiedje, J. M. (2007). DNA–DNA hybridization values and their relationship to whole-genome sequence similarities. *International Journal of Systematic and Evolutionary Microbiology*, **57**, 81–91.

Gravot, A., Dittami, S. M., Rousvoal, S., Lugan, R., Eggert, A., Collén, J., Boyen, C., Bouchereau, A. & Tonon, T. (2010). Diurnal oscillations of metabolite abundances and gene analysis provide new insights into central metabolic processes of the brown alga *Ectocarpus siliculosus*. *New Phytologist*, **188**, 98–110.

Gregg, K. J., Zandberg, W. F., Hehemann, J. H., Whitworth, G. E., Deng, L., Vocadlo, D. J. & Boraston, A. B. (2011). Analysis of a new family of widely distributed metal-independent alpha-mannosidases provides unique insight into the processing of N-linked glycans. *Journal of Biological Chemistry*, **286**, 15586–15596.

Groisillier, A., Labourel, A., Michel, G. & Tonon, T. (2015). The mannitol utilization system of the marine bacterium *Zobellia galactanivorans*. *Applied and Environmental Microbiology*, **81**, 1799–1812.

Grondin, J. M., Tamura, K., Déjean, G., Abbott, D. W. & Brumerb, H. (2017). Polysaccharide utilization loci: fueling microbial communities. *Journal of Bacteriology*, **199**, e00860–16.

Gómez-Pereira, P. R., Fuchs, B. M., Alonso, C., Oliver, M. J., van Beusekom, J. E. E. & Amann, R. (2010). Distinct flavobacterial communities in contrasting water masses of the North Atlantic Ocean. *The ISME Journal*, **4**, 472–487.

Gómez-Pereira, P. R., Hartmann, M., Grob, C., Tarran, G. A., Martin, A. P., Fuchs, B. M., Scanlan, D. J. & Zubkov, M. V. (2013). Comparable light stimulation of organic nutrient uptake by SAR11 and *Prochlorococcus* in the North Atlantic subtropical gyre. *The ISME Journal*, **7**, 603–614.

Gómez-Pereira, P. R., Schüler, M., Fuchs, B. M., Bennke, C., Teeling, H., Waldmann, J., Richter, M., Barbe, V., Bataille, E., Glöckner, F. O. & Amann, R. (2012). Genomic content of uncultured *Bacteroidetes* from contrasting oceanic provinces in the North Atlantic Ocean. *Environmental Microbiology*, **14**, 52–66.

Gügi, B., Le Costaouec, T., Burel, C., Lerouge, P., Helbert, W. & Bardor, M. (2015). Diatom-specific oligosaccharide and polysaccharide structures help to unravel biosynthetic capabilities in diatoms. *Marine Drugs*, **13**, 5993–6018.

Haft, D. H., Selengut, J. D., Richter, R. A., Harkins, D., Basu, M. K. & Beck, E. (2013). TIGRFAMs and genome properties in 2013. *Nucleic Acids Research*, **41**, D387–D395.

Hahnke, R. L., Bennke, C. M., Fuchs, B. M., Mann, A. J., Rhiel, E., Teeling, H., Amann, R. & Harder, J. (2015). Dilution cultivation of marine heterotrophic bacteria abundant after a spring phytoplankton bloom in the North Sea. *Environmental Microbiology*, **17**, 3515–3526.

Hahnke, R. L. & Harder, J. (2013). Phylogenetic diversity of *Flavobacteria* isolated from the North Sea on solid media. *Systematic and Applied Microbiology*, **36**, 497–504.

Halsey, K. H., Carter, A. E. & Giovannoni, S. J. (2012). Synergistic metabolism of a broad range of C1 compounds in the marine methylotrophic bacterium HTCC2181. *Environmental Microbiology*, **14**, 630–640.

Hansell, D. A., Carlson, C. A., Repeta, D. J. & Schlitzer, R. (2009). Dissolved organic matter in the ocean a controversy stimulates new insights. *Oceanography*, **22**, 202–211.

Hecky, R. E., Mopper, K., Kilham, P. & Degens, E. T. (1973). The amino acid and sugar composition of diatom cell-walls. *Marine Biology*, **19**, 323–331.

Hehemann, J.-H., Arevalo, P., Datta, M. S., Yu, X., Corzett, C. H., Henschel, A., Preheim, S. P., Timberlake, S., Alm, E. J. & Polz, M. F. (2016). Adaptive radiation by waves of gene transfer leads to fine-scale resource partitioning in marine microbes. *Nature Communications*, **7**, 12860.

Hehemann, J.-H., Correc, G., Barbeyron, T., Helbert, W., Czjzek, M. & Michel, G. (2010). Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *Nature*, **464**, 908–912.

Hehemann, J.-H., Correc, G., Thomas, F., Bernard, T., Barbeyron, T., Jam, M., Helbert, W., Michel, G. & Czjzek, M. (2012a). Biochemical and structural characterization of the complex agarolytic enzyme system from the marine bacterium *Zobellia galactanivorans*. *Journal of Biological Chemistry*, **287**, 30571–30584.

Hehemann, J.-H., Kelly, A. G., Pudlo, N. A., Martens, E. C. & Boraston, A. B. (2012b). Bacteria of the human gut microbiome catabolize red seaweed glycans with carbohydrate-active enzyme updates from extrinsic microbes. *Proceedings of the National Academy of Sciences of the United States of America*, **109**, 19786–19791.

Hehemann, J.-H., Truong, L. V., Unfried, F., Welsch, N., Kabisch, J., Heiden, S. E., Junker, S., Becher, D., Thürmer, A., Daniel, R., Amann, R. & Schweder, T. (2017). Aquatic adaptation of a laterally acquired pectin degradation pathway in marine gammaproteobacteria. *Environmental Microbiology*, **19**, 2320–2333.

Hellmann, L., Tegel, W., Eggertsson, O., Schweingruber, F. H., Blanchette, R., Kirdyanov, A., Gaertner, H. & Buentgen, U. (2013). Tracing the origin of Arctic driftwood. *Journal of Geophysical Research-Biogeosciences*, **118**, 68–76.

Hemsworth, G. R., Déjean, G., Davies, G. J. & Brumer, H. (2016). Learning from microbial strategies for polysaccharide degradation. *Biochemical Society Transactions*, **44**, 94–108.

Hendry, G. A. F. (1993). Evolutionary origins and natural functions of fructans – a climatological, biogeographic and mechanistic appraisal. *New Phytologist*, **123**, 3–14.

Herlemann, D. P. R., Labrenz, M., Juergens, K., Bertilsson, S., Waniek, J. J. & Andersson, A. F. (2011). Transitions in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *The ISME Journal*, **5**, 1571–1579.

Hoagland, K. D., Rosowski, J. R., Gretz, M. R. & Roemer, S. C. (1993). Diatom extracellular polymeric substances: function, fine structure, chemistry, and physiology. *Journal of Phycology*, **29**, 537 – 566.

Hug, L. A., Baker, B. J., Anantharaman, K., Brown, C. T., Probst, A. J., Castelle, C. J., Butterfield, C. N., Hernsdorf, A. W., Amano, Y., Ise, K., Suzuki, Y., Dudek, N., Relman, D. A., Finstad, K. M., Amundson, R., Thomas, B. C. & Banfield, J. F. (2016). A new view of the tree of life. *Nature Microbiology*, **1**, 16048.

Hugerth, L. W., Larsson, J., Alneberg, J., Lindh, M. V., Legrand, C., Pinhassi, J. & Andersson, A. F. (2015). Metagenome-assembled genomes uncover a global brackish microbiome. *Genome Biology*, **16**, 279.

Huntemann, M., Ivanova, N. N., Mavromatis, K., Tripp, H. J., Paez-Espino, D., Palaniappan, K., Szeto, E., Pillay, M., Chen, I. M. A., Pati, A., Nielsen, T., Markowitz, V. M. & Kyrpides, N. C. (2015). The standard operating procedure of the DOE-JGI Microbial Genome Annotation Pipeline (MGAP v.4). *Standards in Genomic Sciences*, **10**, 86.

Hutchinson, G. (1961). The paradox of the plankton. *American Naturalist*, **95**, 137–145.

Hyatt, D., Chen, G.-L., LoCascio, P. F., Land, M. L., Larimer, F. W. & Hauser, L. J. (2010). Prodigal: prokaryotic gene recognition and translation initiation site identification. *BMC Bioinformatics*, **11**, 119.

Imelfort, M., Parks, D., Woodcroft, B. J., Dennis, P., Hugenholtz, P. & Tyson, G. W. (2014). GroopM: an automated tool for the recovery of population genomes from related metagenomes. *PeerJ*, **2**, e603.

Jiao, N., Herndl, G. J., Hansell, D. A., Benner, R., Kattner, G., Wilhelm, S. W., Kirchman, D. L., Weinbauer, M. G., Luo, T., Chen, F. & Azam, F. (2010). Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. *Nature Reviews Microbiology*, **8**, 593–599.

Kabisch, A., Otto, A., König, S., Becher, D., Albrecht, D., Schüler, M., Teeling, H., Amann, R. I. & Schweder, T. (2014). Functional characterization of polysaccharide utilization loci in the marine *Bacteroidetes* '*Gramella forsetii*' KT0803. *The ISME Journal*, **8**, 1492–1502.

Kang, D. D., Froula, J., Egan, R. & Wang, Z. (2015). MetaBAT, an efficient tool for accurately reconstructing single genomes from complex microbial communities. *PeerJ*, **3**, e1165.

Kappelmann, L., Krüger, K., Hehemann, J.-H., Harder, J., Markert, S., Unfried, F., Becher, D., Shapiro, N., Schweder, T., Amann, R. I. & Teeling, H. (2018). Polysaccharide utilization loci of North Sea *Flavobacteriia* as basis for using SusC/D-protein expression for predicting major phytoplankton glycans. *The ISME Journal*, **13**, 76–91.

Karp, P. D., Latendresse, M., Paley, S. M., Krummenacker, M., Ong, Q. D., Billington, R., Kothari, A., Weaver, D., Lee, T., Subhraveti, P., Spaulding, A., Fulcher, C., Keseler, I. M. & Caspi, R. (2016). Pathway Tools version 19.0 update: software for pathway/genome informatics and systems biology. *Briefings in Bioinformatics*, **17**, 877–890.

Katoh, K. & Standley, D. M. (2013). MAFFT multiple sequence alignment software version 7: improvements in performance and usability. *Molecular Biology and Evolution*, **30**, 772–780.

Keith, S. C. & Arnosti, C. (2001). Extracellular enzyme activity in a river-bay-shelf transect: variations in polysaccharide hydrolysis rates with substrate and size class. *Aquatic Microbial Ecology*, **24**, 243–253.

Khan, S. T., Nakagawa, Y. & Harayama, S. (2007). *Sediminibacter furfurosus* gen. nov., sp nov and *Gilvibacter sediminis* gen. nov., sp nov., novel members of the family *Flavobacteriaceae*. *International Journal of Systematic and Evolutionary Microbiology*, **57**, 265–269.

Kirchman, D. L. (2002). The ecology of *Cytophaga-Flavobacteria* in aquatic environments. *FEMS Microbiology Ecology*, **39**, 91–100.

Kitamura, M., Okuyama, M., Tanzawa, F., Mori, H., Kitago, Y., Watanabe, N., Kimura, A., Tanaka, I. & Yao, M. (2008). Structural and functional analysis of a glycoside hydrolase family 97 enzyme from *Bacteroides thetaiotaomicron*. *Journal of Biological Chemistry*, **283**, 36328–36337.

Klindworth, A., Mann, A. J., Huang, S., Wichels, A., Quast, C., Waldmann, J., Teeling, H. & Glöckner, F. O. (2014). Diversity and activity of marine bacterioplankton during a diatom bloom in the North Sea assessed by total RNA and pyrotag sequencing. *Marine Genomics*, **18**, 185–192.

Kloareg, B. & Quatrano, R. S. (1988). Structure of the cell walls of marine algae and ecophysiological functions of the matrix polysaccharides. *Oceanography and Marine Biology Annual Review*, **26**, 259–315.

Koch, H., Durwald, A., Schweder, T., Noriega-Ortega, B., Vidal-Melgosa, S., Hehemann, J. H., Dittmar, T., Freese, H. M., Becher, D., Simon,

M. & Wietz, M. (2018). Biphasic cellular adaptations and ecological implications of *Alteromonas macleodii* degrading a mixture of algal polysaccharides. *The ISME Journal*, **13**, 92–103.

Kolton, M., Sela, N., Elad, Y. & Cytryn, E. (2013). Comparative genomic analysis indicates that niche adaptation of terrestrial *Flavobacteria* is strongly linked to plant glycan metabolism. *PLoS ONE*, **8**, e76704.

Konstantinidis, K. T. & Rossello-Mora, R. (2015). Classifying the uncultivated microbial majority: a place for metagenomic data in the *Candidatus* proposal. *Systematic and Applied Microbiology*, **38**, 223–230.

Konstantinidis, K. T., Rosselló-Móra, R. & Amann, R. (2017). Uncultivated microbes in need of their own taxonomy. *The ISME Journal*, **11**, 2399–2406.

Konstantinidis, K. T. & Tiedje, J. M. (2005). Genomic insights that advance the species definition for prokaryotes. *Proceedings of the National Academy of Sciences of the United States of America*, **102**, 2567–2572.

Koropatkin, N. M., Martens, E. C., Gordon, J. I. & Smith, T. J. (2008). Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices. *Structure*, **16**, 1105–1115.

Kraan, S. (2012). Algal polysaccharides, novel applications and outlook. In C.-F. Chang, ed., *Carbohydrates - comprehensive studies on glycobiology and glycotechnology*, 489–532. InTech.

Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., Edwards, R. A. & Stoye, J. (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Research*, **36**, 2230–2239.

Krogh, A., Larsson, B., von Heijne, G. & Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *Journal of Molecular Biology*, **305**, 567–580.

Kuhaudomlarp, S., Patron, N. J., Henrissat, B., Rejzek, M., Saalbach, G. & Field, R. A. (2018). Identification of *Euglena gracilis* β-1,3-glucan phosphorylase and establishment of a new glycoside hydrolase (GH) family GH149. *Journal of Biological Chemistry*, **293**, 2865–2876.

Kumar, S., Tamura, K. & Nei, M. (2004). MEGA3: integrated software for molecular evolutionary genetics analysis and sequence alignment. *Briefings in Bioinformatics*, **5**, 150–163.

Labourel, A., Jam, M., Jeudy, A., Hehemann, J.-H., Czjzek, M. & Michel, G. (2014). The beta-glucanase ZgLamA from *Zobellia galactanivorans* evolved a bent active site adapted for efficient degradation of algal laminarin. *Journal of Biological Chemistry*, **289**, 2027–2042.

Labourel, A., Jam, M., Legentil, L., Sylla, B., Hehemann, J.-H., Ferrieres, V., Czjzek, M. & Michel, G. (2015). Structural and biochemical characterization of the laminarinase ZgLamC(GH16) from *Zobellia galactanivorans* suggests preferred recognition of branched laminarin. *Acta Crystallographica Section D Structural Biology*, **71**, 173–184.

Lagkouvardos, I., Joseph, D., Kapfhammer, M., Giritli, S., Horn, M., Haller, D. & Clavel, T. (2016). IMNGS: A comprehensive open resource of processed 16S rRNA microbial profiles for ecology and diversity studies. *Scientific Reports*, **6**, 33721.

Lahaye, M. & Robic, A. (2007). Structure and functional properties of ulvan, a polysaccharide from green seaweeds. *Biomacromolecules*, **8**, 1765–1774.

Laine, R. A. (1994). A calculation of all possible oligosaccharide isomers both branched and linear yields $1.05\text{x}10^{12}$ structures for a reducing hexasaccharide - the isomer-barrier to development of single-method saccharide sequencing or synthesis systems. *Glycobiology*, **4**, 759–767.

Landa, M., Blain, S., Christaki, U., Monchy, S. & Obernosterer, I. (2016). Shifts in bacterial community composition associated with increased carbon cycling in a mosaic of phytoplankton blooms. *The ISME Journal*, **10**, 39–50.

Larsbrink, J., Rogers, T. E., Hemsworth, G. R., McKee, L. S., Tauzin, A. S., Spadiut, O., Klinter, S., Pudlo, N. A., Urs, K., Koropatkin, N. M., Creagh, A. L., Haynes, C. A., Kelly, A. G., Cederholm, S. N., Davies, G. J., Martens, E. C. & Brumer, H. (2014). A discrete genetic locus confers xyloglucan metabolism in select human gut *Bacteroidetes*. *Nature*, **506**, 498–502.

Le Costaouëc, T., Unamunzaga, C., Mantecon, L. & Helbert, W. (2017). New structural insights into the cell-wall polysaccharide of the diatom *Phaeodactylum tricornutum*. *Algal Research*, **26**, 172–179.

Lee, S. C., Gepts, P. L. & Whitaker, J. R. (2002). Protein structures of common bean (*Phaseolus vulgaris*) alpha-amylase inhibitors. *Journal of Agricultural and Food Chemistry*, **50**, 6618–6627.

Letunic, I. & Bork, P. (2016). Interactive tree of life (iTOL) v3: an online tool for the display and annotation of phylogenetic and other trees. *Nucleic Acids Research*, **44**, W242–W245.

Li, B., Lu, F., Wei, X. & Zhao, R. (2008). Fucoidan: structure and bioactivity. *Molecules*, **13**, 1671–1695.

Li, D., Liu, C. M., Luo, R., Sadakane, K. & Lam, T. W. (2015a). MEGAHIT: an ultra-fast single-node solution for large and complex metagenomics assembly via succinct de Bruijn graph. *Bioinformatics*, **31**, 1674–1676.

Li, H., Handsaker, B., Wysoker, A., Fennell, T., Ruan, J., Homer, N., Marth, G., Abecasis, G., Durbin, R. & Subgroup, . G. P. D. P. (2009). The Sequence Alignment/Map format and SAMtools. *Bioinformatics*, **25**, 2078–2079.

Li, J., Wei, B., Wang, J., Liu, Y., Dasgupta, S., Zhang, L. & Fang, J. (2015b). Variation in abundance and community structure of particle-attached and free-living bacteria in the South China Sea. *Deep-Sea Research II*, **122**, 64–73.

Lindh, M. V., Sjöstedt, J., Andersson, A. F., Baltar, F., Hugerth, L. W., Lundin, D., Muthusamy, S., Legrand, C. & Pinhassi, J. (2015). Disentangling seasonal bacterioplankton population dynamics by high-frequency sampling. *Environmental Microbiology*, **17**, 2459–2476.

Liu, N., Li, H., Chevrette, M. G., Zhang, L., Cao, L., Zhou, H., Zhou, X., Zhou, Z., Pope, P. B., Currie, C. R., Huang, Y. & Wang, Q. (2018). Functional metagenomics reveals abundant polysaccharide-degrading gene clusters and cellobiose utilization pathways within gut microbiota of a wood-feeding higher termite. *The ISME Journal*, **13**, 104–117.

Loebl, M., van Beusekom, J. E. E. & Philippart, C. J. M. (2013). No microzooplankton grazing during a *Mediopyxis helysia* dominated diatom bloom. *Journal of Sea Research*, **82**, 80–85.

Lombard, V., Bernard, T., Rancurel, C., Brumer, H., Coutinho, P. M. & Henrissat, B. (2010). A hierarchical classification of polysaccharide lyases for glycogenomics. *Biochemical Journal*, **432**, 437–444.

Lombard, V., Golaconda Ramulu, H., Drula, E., Coutinho, P. M. & Henrissat, B. (2014). The carbohydrate-active enzymes database (CAZy) in 2013. *Nucleic Acids Research*, **42**, D490–D495.

Longhurst, A. R. (2006). *Ecological geography of the sea.* Academic Press.

Lucas, J., Wichels, A., Teeling, H., Chafee, M., Scharfe, M. & Gerdts, G. (2015). Annual dynamics of North Sea bacterioplankton: seasonal variability superimposes short-term variation. *FEMS Microbiology Ecology*, **91**, fiv099.

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

Luo, R., Liu, B., Xie, Y., Li, Z., Huang, W. et al. (2012). SOAPdenovo2: an empirically improved memory-efficient short-read de novo assembler. *Gigascience*, **1**, 18.

Malinsky-Rushansky, N. Z. & Legrand, C. (1996). Excretion of dissolved organic carbon by phytoplankton of different sizes and subsequent bacterial uptake. *Marine Ecology Progress Series*, **132**, 249–255.

Mann, A. J., Hahnke, R. L., Huang, S., Werner, J., Xing, P., Barbeyron, T., Huettel, B., Stüber, K., Reinhardt, R., Harder, J., Glöckner, F. O., Amann, R. I. & Teeling, H. (2013). The genome of the alga-associated marine flavobacterium *Formosa agariphila* KMM 3901^{T} reveals a broad potential for degradation of algal polysaccharides. *Applied and Environmental Microbiology*, **79**, 6813–6822.

Mann, D. G. (1999). The species concept in diatoms. *Phycologia*, **38**, 437–495.

Markowitz, V. M., Chen, I. M. A., Chu, K., Szeto, E., Palaniappan, K. et al. (2014). IMG/M 4 version of the integrated metagenome comparative analysis system. *Nucleic Acids Research*, **42**, D568–D573.

Martens, E. C., Kelly, A. G., Tauzin, A. S. & Brumer, H. (2014). The devil lies in the details: how variations in polysaccharide fine-structure impact the physiology and evolution of gut microbes. *Journal of Molecular Biology*, **426**, 3851–3865.

Martens, E. C., Koropatkin, N. M., Smith, T. J. & Gordon, J. I. (2009). Complex glycan catabolism by the human gut microbiota: the *Bacteroidetes* Sus-like paradigm. *Journal of Biological Chemistry*, **284**, 24673–24677.

Martens, E. C., Lowe, E. C., Chiang, H., Pudlo, N. A., Wu, M., McNulty, N. P., Abbott, D. W., Henrissat, B., Gilbert, H. J., Bolam, D. N. & Gordon, J. I. (2011). Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biology*, **9**, e1001221.

Martin, M., Portetelle, D., Michel, G. & Vandenbol, M. (2014). Microorganisms living on macroalgae: diversity, interactions, and biotechnological applications. *Applied Microbiology and Biotechnology*, **98**, 2917–2935.

Martinez-Garcia, M., Brazel, D. M., Swan, B. K., Arnosti, C., Chain, P. S. G. et al. (2012). Capturing single cell genomes of active polysaccharide degraders: an unexpected contribution of *Verrucomicrobia*. *PLoS ONE*, **7**, e35314.

McCarthy, A. (2010). Third generation DNA sequencing: pacific biosciences' single molecule real time technology. *Chemistry & Biology*, **17**, 675–676.

McCoy, A. J., Grosse-Kunstleve, R. W., Adams, P. D., Winn, M. D., Storoni, L. C. & Read, R. J. (2007). Phaser crystallographic software. *Journal of Applied Crystallography*, **40**, 658–674.

Mende, D. R., Sunagawa, S., Zeller, G. & Bork, P. (2013). Accurate and universal delineation of prokaryotic species. *Nature Methods*, **10**, 881–884.

Mewis, K., Lenfant, N., Lombard, V. & Henrissat, B. (2016). Dividing the large glycoside hydrolase family 43 into subfamilies: a motivation for detailed enzyme characterization. *Applied and Environmental Microbiology*, **82**, 1686–1692.

Meyer, F., Goesmann, A., McHardy, A. C., Bartels, D., Bekel, T., Clausen, J., Kalinowski, J., Linke, B., Rupp, O., Giegerich, R. & Puhler, A. (2003). GenDB - an open source genome annotation system for prokaryote genomes. *Nucleic Acids Research*, **31**, 2187–2195.

Mhanna, R., Kashyap, A., Palazzolo, G., Vallmajo-Martin, Q., Becher, J., Möller, S., Schnabelrauch, M. & Zenobi-Wong, M. (2014). Chondrocyte culture in three dimensional alginate sulfate hydrogels promotes proliferation while maintaining expression of chondrogenic markers. *Tissue Engineering: Part A*, **20**, 1454–1464.

Michel, G., Barbeyron, T., Kloareg, B. & Czjzek, M. (2009). The family 6 carbohydrate-binding modules have coevolved with their appended catalytic modules toward similar substrate specificity. *Glycobiology*, **19**, 615–623.

Michel, G., Chantalat, L., Duee, E., Barbeyron, T., Henrissat, B., Kloareg, B. & Dideberg, O. (2001). The kappa-carrageenase of *P. carrageenovora* features a tunnel-shaped active site: a novel insight in the evolution of Clan-B glycoside hydrolases. *Structure*, **9**, 513–525.

Michel, G., Tonon, T., Scornet, D., Cock, J. M. & Kloareg, B. (2010). Central and storage carbon metabolism of the brown alga *Ectocarpus siliculosus*: insights into the origin and evolution of storage carbohydrates in Eukaryotes. *New Phytologist*, **188**, 67–81.

Mikheenko, A., Valin, G., Prjibelski, A., Saveliev, V. & Gurevich, A. (2016). Icarus: visualizer for *de novo* assembly evaluation. *Bioinformatics*, **32**, 3321–3323.

Mikheyev, A. S. & Tin, M. M. Y. (2014). A first look at the Oxford Nanopore MinION sequencer. *Molecular Ecology Resourses*, **14**, 1097–1102.

Moran, M. A., Belas, R., Schell, M. A., Gonzalez, J. M., Sun, F. et al. (2007). Ecological genomics of marine Roseobacters. *Applied and Environmental Microbiology*, **73**, 4559–4569.

Moreira, L. R. S. & Filho, E. X. F. (2008). An overview of mannan structure and mannan-degrading enzyme systems. *Applied Microbiology and Biotechnology*, **79**, 165–178.

Morris, R. M., Nunn, B. L., Frazar, C., Goodlett, D. R., Ting, Y. S. & Rocap, G. (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *The ISME Journal*, **4**, 673–685.

Murray, A. E., Arnosti, C., De La Rocha, C. L., Grossart, H.-P. & Passow, U. (2007). Microbial dynamics in autotrophic and heterotrophic seawater mesocosms. II. Bacterioplankton community structure and hydrolytic enzyme activities. *Aquatic Microbial Ecology*, **49**, 123–141.

Murshudov, G. N., Skubák, P., Lebedev, A. A., Pannu, N. S., Steiner, R. A., Nicholls, R. A., Winn, M. D., Long, F. & Vagin, A. A. (2011). REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallographica Section D Biological Crystallography*, **67**, 355–367.

Muñoz-Marín, M. d. C., Luque, I., Zubkov, M. V., Hill, P. G., Diez, J. & García-Fernández, J. M. (2013). *Prochlorococcus* can use the Pro1404 transporter to take up glucose at nanomolar concentrations in the Atlantic Ocean. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 8597–8602.

Myklestad, S. (1974). Production of carbohydrates by marine planktonic diatoms. I. Comparison of nine different species in culture. *Journal of Experimental marine Biology and Ecology*, **15**, 261–274.

Myklestad, S. M. (1989). Production, chemical structure, metabolism, and biological function of the (1→3)-linked, β3-D-glucans in diatoms. *Biological Oceanography*, **6**, 313–326.

Myklestad, S. M. & Granum, E. (2009). Biology of (1,3)-beta-glucans and related glucans in Protozoans and Chromistans. In A. Bacic, G. Fincher & B. Stone, eds., *Chemistry, biochemistry, and biology of 1-3 beta glucans and related polysaccharides*, 353–385. Academic Press.

Mystkowska, A. A., Robb, C., Vidal-Melgosa, S., Vanni, C., Fernandez-Guerra, A., Hohne, M. & Hehemann, J. H. (2018). Molecular recognition of the beta-glucans laminarin and pustulan by a SusD-like glycan-binding protein of a marine *Bacteroidetes*. *FEBS Journal*, **285**, 4465–4481.

Mühlenbruch, M., Grossart, H. P., Eigemann, F. & Voss, M. (2018). Mini-review: Phytoplankton-derived polysaccharides in the marine environment and their interactions with heterotrophic bacteria. *Environmental Microbiology*, **20**, 2671–2685.

Naumoff, D. G. & Dedysh, S. N. (2012). Lateral gene transfer between the *Bacteroidetes* and Acidobacteria: the case of alpha-L-rhamnosidases. *FEBS Letters*, **586**, 3843–3851.

Ndeh, D., Rogowski, A., Cartmell, A., Luis, A. S., Baslé, A. et al. (2017). Complex pectin metabolism by gut bacteria reveals novel catalytic functions. *Nature*, **544**, 65–70.

Nedashkovskaya, O. I., Kim, S. B., Han, S. K., Rhee, M. S., Lysenko, A. M., Falsen, E., Frolova, G. M., Mikhailov, V. V. & Bae, K. S. (2004). *Ulvibacter litoralis* gen. nov., sp nov., a novel member of the family *Flavobacteriaceae* isolated from the green alga *Ulva fenestrata*. *International Journal of Systematic and Evolutionary Microbiology*, **54**, 119–123.

Nedashkovskaya, O. I., Kim, S. B., Vancanneyt, M., Snauwaert, C., Lysenko, A. M., Rohde, M., Frolova, G. M., Zhukova, N. V., Mikhailov, V. V., Bae, K. S., Oh, H. W. & Swings, J. (2006). *Formosa agariphila* sp nov., a budding bacterium of the family *Flavobacteriaceae* isolated from marine environments, and emended description of the genus *Formosa*. *International Journal of Systematic and Evolutionary Microbiology*, **56**, 161–167.

Needham, D. M., Fichot, E. B., Wang, E., Berdjeb, L., Cram, J. A., Fichot, C. G. & Fuhrman, J. A. (2018). Dynamics and interactions of highly resolved marine plankton via automated high-frequency sampling. *The ISME Journal*, **12**, 2417–2432.

Needham, D. M. & Fuhrman, J. A. (2016). Pronounced daily succession of phytoplankton, archaea and bacteria following a spring bloom. *Nature Microbiology*, **1**, 16005.

Nelson, D. M., Treguer, P., Brzezinski, M. A., Leynaert, A. & Queguiner, B. (1995). Production and dissolution of biogenic silica in the ocean - revised global estimates, comparison with regional data and relationship to biogenic sedimentation. *Global Biogeochemical Cycles*, **9**, 359–372.

Neumann, A. M., Balmonte, J. P., Berger, M., Giebel, H.-A., Arnosti, C., Voget, S., Simon, M., Brinkhoff, T. & Wietz, M. (2015). Different utilization of alginate and other algal polysaccharides by marine *Alteromonas macleodii* ecotypes. *Environmental Microbiology*, **17**, 3857–3868.

Newton, R. J., Griffin, L. E., Bowles, K. M., Meile, C., Gifford, S., Givens, C. E., Howard, E. C., King, E., Oakley, C. A., Reisch, C. R., Rinta-Kanto, J. M., Sharma, S., Sun, S., Varaljay, V., Vila-Costa, M., Westrich, J. R. & Moran, M. A. (2010). Genome characteristics of a generalist marine bacterial lineage. *The ISME Journal*, **4**, 784–798.

Nielsen, H., Brunak, S. & von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. *Protein Engineering*, **12**, 3–9.

Niu, Y., Shen, H., Chen, J., Xie, P., Yang, X., Tao, M., Ma, Z. & Qi, M. (2011). Phytoplankton community succession shaping bacterioplankton community composition in Lake Taihu, China. *Water Research*, **45**, 4169–4182.

Nurk, S., Meleshko, D., Korobeynikov, A. & Pevzner, P. A. (2017). metaSPAdes: a new versatile metagenomic assembler. *Genome Research*, **27**, 824–834.

Oh, H.-M., Giovannoni, S. J., Lee, K., Ferriera, S., Johnson, J. & Cho, J.-C. (2009). Complete genome sequence of *Robiginitalea biformata* HTCC2501. *Journal of Bacteriology*, **191**, 7144–7145.

Okuda, K. (2002). Structure and phylogeny of cell coverings. *Journal of Plant Research*, **115**, 283–288.

Olenina, I., Hajdu, S., Edler, L., Andersson, A., Wasmund, N., Busch, S., Göbel, J., Gromisz, S., Huseby, S., Huttunen, M., Jaanus, A., Kokkonen, P., Ledaine, I. & Niemkiewicz, E. (2006). Biovolumes and size-classes of phytoplankton in the Baltic Sea. *HELCOM Baltic Sea Environment Proceedings*, **106**, 144pp.

Ondov, B. D., Treangen, T. J., Melsted, P., Mallonee, A. B., Bergman, N. H., Koren, S. & Phillippy, A. M. (2016). Mash: fast genome and metagenome distance estimation using MinHash. *Genome Biology*, **17**, 132.

Osterholz, H., Niggemann, J., Giebel, H. A., Simon, M. & Dittmar, T. (2015). Inefficient microbial production of refractory dissolved organic matter in the ocean. *Nature Communications*, **6**, 7422.

Otto, A., Bernhardt, J., Meyer, H., Schaffer, M., Herbst, F. A., Siebourg, J., Mäder, U., Lalk, M., Hecker, M. & Becher, D. (2010). Systems-wide temporal proteomic profiling in glucose-starved *Bacillus subtilis*. *Nature Communications*, **1**, 137.

Painter, T. J. (1983). Algal Polysaccharides. In G. O. Aspinall, ed., *The Polysaccharides*, 195–285. Academic Press.

Panschin, I., Huang, S., Meier-Kolthoff, J. P., Tindall, B. J., Rohde, M. et al. (2016). Comparing polysaccharide decomposition between the type strains *Gramella echinicola* KMM 6050^{T} (DSM 19838^{T}) and *Gramella portivictoriae* UST040801-001^{T} (DSM 23547^{T}), and emended description of *Gramella echinicola* Nedashkovskaya et al. 2005 emend. Shahina et al. 2014 and *Gramella portivictoriae* Lau et al. 2005. *Standards in Genomic Sciences*, **11**, 37.

Parks, D. H., Chuvochina, M., Waite, D. W., Rinke, C., Skarshewski, A., Chaumeil, P.-A. & Hugenholtz, P. (2018). A standardized bacterial taxonomy based on genome phylogeny substantially revises the tree of life. *Nature Biotechnology*, **36**, 996–1004.

Parks, D. H., Imelfort, M., Skennerton, C. T., Hugenholtz, P. & Tyson, G. W. (2015). CheckM: assessing the quality of microbial genomes recovered from isolates, single cells, and metagenomes. *Genome Research*, **25**, 1043–1055.

Parks, D. H., Rinke, C., Chuvochina, M., Chaumeil, P.-A., Woodcroft, B. J., Evans, P. N., Hugenholtz, P. & Tyson, G. W. (2017). Recovery of nearly 8,000 metagenome-assembled genomes substantially expands the tree of life. *Nature Microbiology*, **2**, 1533–1542.

Passow, U. (2002). Transparent exopolymer particles (TEP) in aquatic environments. *Progress in Oceanography*, **55**, 287–333.

Peng, Y., Leung, H. C., Yiu, S. M. & Chin, F. Y. (2012). IDBA-UD: a de novo assembler for single-cell and metagenomic sequencing data with highly uneven depth. *Bioinformatics*, **28**, 1420–1428.

Pernthaler, J. (2005). Predation on prokaryotes in the water column and its ecological implications. *Nature Reviews Microbiology*, **3**, 537–546.

Podell, S. & Gaasterland, T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biology*, **8**, R16.

Popper, Z. A., Michel, G., Hervé, C., Domozych, D. S., Willats, W. G. T., Tuohy, M. G., Kloareg, B. & Stengel, D. B. (2011). Evolution and diversity of plant cell walls: from algae to flowering plants. *Annual Review of Plant Biology*, **62**, 567–590.

Preiss, J. (1984). Bacterial glycogen synthesis and its regulation. *Annual Review of Microbiology*, **38**, 419–458.

Price, M. N., Dehal, P. S. & Arkin, A. P. (2010). FastTree 2–approximately maximum-likelihood trees for large alignments. *PLoS ONE*, **5**, e9490.

Pruesse, E., Peplies, J. & Glöckner, F. O. (2012). SINA: Accurate high-throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, **28**, 1823–1829.

Punta, M., Coggill, P. C., Eberhardt, R. Y., Mistry, J., Tate, J., Boursnell, C., Pang, N., Forslund, K., Ceric, G., Clements, J., Heger, A., Holm, L., Sonnhammer, E. L., Eddy, S. R., Bateman, A. & Finn, R. D. (2012). The Pfam protein families database. *Nucleic Acids Research*, **40**, D290–D301.

Qin, Z., Yan, Q., Lei, J., Yang, S., Jiang, Z. & Wu, S. (2015). The first crystal structure of a glycoside hydrolase family 17 β-1,3-glucanosyltransferase displays a unique catalytic cleft. *Acta Crystallographica Section D Biological Crystallography*, **71**, 1714–1724.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J. & Glöckner, F. O. (2013). The SILVA ribosomal RNA gene database project: improved data processing and web-based tools. *Nucleic Acids Research*, **41**, D590–D596.

Quince, C., Delmont, T. O., Raguideau, S., Alneberg, J., Darling, A. E., Collins, G. & Eren, A. M. (2017). DESMAN: a new tool for de novo extraction of strains from metagenomes. *Genome Biology*, **18**, 181.

Rakoff-Nahoum, S., Foster, K. R. & Comstock, L. E. (2016). The evolution of cooperation within the gut microbiota. *Nature*, **533**, 255–259.

Rawlings, N. D., Barrett, A. J. & Bateman, A. (2012). MEROPS: the database of proteolytic enzymes, their substrates and inhibitors. *Nucleic Acids Research*, **40**, D343–D350.

Read, S. M., Currie, G. & Bacic, A. (1996). Analysis of the structural heterogeneity of laminarin by electrospray-ionisation-mass spectrometry. *Carbohydrate Research*, **281**, 187–201.

Reddy, T. B. K., Thomas, A. D., Stamatis, D., Bertsch, J., Isbandi, M., Jansson, J., Mallajosyula, J., Pagani, I., Lobos, E. A. & Kyrpides, N. C. (2015). The Genomes OnLine Database (GOLD) v.5: a metadata management system based on a four level (meta)genome project classification. *Nucleic Acids Research*, **43**, D1099–D1106.

Rees, D. A. & Welsh, E. J. (1977). Secondary and tertiary structure of polysaccharides in solutions and gels. *Angewandte Chemie International Edition*, **16**, 214–224.

Reese, E. T. & Mandels, M. (1959). Beta-D-1,3 glucanases in fungi. *Canadian Journal of Microbiology*, **5**, 173–185.

Reeves, A. R., Wang, G. R. & Salyers, A. A. (1997). Characterization of four outer membrane proteins that play a role in utilization of starch by *Bacteroides thetaiotaomicron*. *Journal of Bacteriology*, **179**, 643–649.

Reintjes, G., Arnosti, C., Fuchs, B. & Amann, R. (2018). Selfish, sharing and scavenging bacteria in the Atlantic Ocean: a biogeographical study of bacterial substrate utilisation. *The ISME Journal*, https://doi.org/10.1038/s41396–018–0326–3.

Reintjes, G., Arnosti, C., Fuchs, B. M. & Amann, R. (2017). An alternative polysaccharide uptake mechanism of marine bacteria. *The ISME Journal*, **11**, 1640–1650.

Reisky, L., Stanetty, C., Mihovilovic, M. D., Schweder, T., Hehemann, J. H. & Bornscheuer, U. T. (2018). Biochemical characterization of an ulvan lyase from the marine flavobacterium *Formosa agariphila* KMM 3901^{T}. *Applied Microbiology Biotechnology*, **102**, 6987–6996.

Rinta-Kanto, J. M., Sun, S., Sharma, S., Kiene, R. P. & Moran, M. A. (2012). Bacterial community transcription patterns during a marine phytoplankton bloom. *Environmental Microbiology*, **14**, 228–239.

Rodriguez-R, L. M. & Konstantinidis, K. T. (2016). The enveomics collection: a toolbox for specialized analyses of microbial genomes and metagenomes. *PeerJ Preprints*, **4**, e1900v1.

Romine, M. F. (2011). Genome-wide protein localization prediction strategies for gram negative bacteria. *BMC Genomics*, **12**, S1.

Ruff, S. E., Probandt, D., Zinkann, A.-C., Iversen, M. H., Klaas, C., Würzberg, L., Krombholz, N., Wolf-Gladrow, D., Amann, R. & Knittel, K. (2014). Indications for algae-degrading benthic microbial communities in deep-sea sediments along the Antarctic Polar Front. *Deep-Sea Research II*, **108**, 6–16.

Salinas, A. & French, C. E. (2017). The enzymatic ulvan depolymerisation system from the alga-associated marine flavobacterium *Formosa agariphila*. *Algal Research*, **27**, 335–344.

Sanger, F., Nicklen, S. & Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. *Proceedings of the National Academy of Sciences of the United States of America*, **74**, 5463–5467.

Sapp, M., Wichels, A., Wiltshire, K. H. & Gerdts, G. (2007). Bacterial community dynamics during the winter-spring transition in the North Sea. *FEMS Microbiology Ecology*, **59**, 622–637.

Sarmento, H. & Gasol, J. M. (2012). Use of phytoplankton-derived dissolved organic carbon by different types of bacterioplankton. *Environmental Microbiology*, **14**, 2348–2360.

Sarmento, H., Morana, C. & Gasol, J. M. (2016). Bacterioplankton niche partitioning in the use of phytoplankton-derived dissolved organic carbon: quantity is more important than quality. *The ISME Journal*, **10**, 2582–2592.

Schattenhofer, M., Fuchs, B. M., Amann, R., Zubkov, M. V., Tarran, G. A. & Pernthaler, J. (2009). Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environmental Microbiology*, **11**, 2078–2093.

Seemann, T. (2014). Prokka: rapid prokaryotic genome annotation. *Bioinformatics*, **30**, 2068–2069.

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., Richter, A. R. & White, O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. *Nucleic Acids Research*, **35**, D260–D264.

Senoura, T., Ito, S., Taguchi, H., Higa, M., Hamada, S., Matsui, H., Ozawa, T., Jin, S., Watanabe, J., Wasaki, J. & Ito, S. (2011). New microbial mannan catabolic pathway that involves a novel mannosylglucose phosphorylase. *Biochemical and Biophysical Research Communications*, **408**, 701–706.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B. & Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. *Genome Research*, **13**, 2498–2504.

Shin, J. B., Krey, J. F., Hassan, A., Metlagel, Z., Tauscher, A. N., Pagana, J., Sherman, N. E., Jeffery, E. D., Spinelli, K. J., Zhao, H., Wilmarth, P. A., Choi, D., David, L. L., Auer, M. & Barr-Gillespie, P. G. (2013). Molecular architecture of the chick vestibular hair bundle. *Nature Neuroscience*, **16**, 365–374.

Shipman, J. A., Berleman, J. E. & Salyers, A. A. (2000). Characterization of four outer membrane proteins involved in binding starch to the cell surface of *Bacteroides thetaiotaomicron*. *Journal of Bacteriology*, **182**, 5365–5372.

Simpson, J. T. & Pop, M. (2015). The theory and practice of genome sequence assembly. *Annual Review of Genomics and Human Genetics*, **16**, 153–172.

Smith, K. A. & Salyers, A. A. (1991). Characterization of a neopullulanase and an alpha-glucosidase from *Bacteroides thetaiotaomicron* 95-1. *Journal of Bacteriology*, **173**, 2962–2968.

Sonnenburg, E. D., Zheng, H., Joglekar, P., Higginbottom, S. K., Firbank, S. J., Bolam, D. N. & Sonnenburg, J. L. (2010). Specificity of polysaccharide use in intestinal *Bacteroides* species determines diet-induced microbiota alterations. *Cell*, **141**, 1241–1252.

Soo, R. M., Skennerton, C. T., Sekiguchi, Y., Imelfort, M., Paech, S. J., Dennis, P. G., Steen, J. A., Parks, D. H., Tyson, G. W. & Hugenholtz, P. (2014). An expanded genomic representation of the phylum *Cyanobacteria*. *Genome Biology and Evolution*, **6**, 1031–1045.

Sperling, M., Piontek, J., Engel, A., Wiltshire, K. H., Niggemann, J., Gerdts, G. & Wichels, A. (2017). Combined carbohydrates support rich communities of particle-associated marine bacterioplankton. *Frontiers in Microbiology*, **8**, 65.

St John, F. J., González, J. M. & Pozharski, E. (2010). Consolidation of glycosyl hydrolase family 30: a dual domain 4/7 hydrolase family consisting of two structurally distinct groups. *FEBS Letters*, **584**, 4435–4441.

Stam, M. R., Danchin, E. G. J., Rancurel, C., Coutinho, P. M. & Henrissat, B. (2006). Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Engineering Design and Selection*, **19**, 555–562.

Stamatakis, A. (2014). RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics*, **30**, 1312–1313.

Stanier, R. Y. (1947). Studies on nonfruiting myxobacteria: I. *Cytophaga johnsonae*, n. sp., a chitin-decomposing myxobacterium. *Journal of Bacteriology*, **53**, 297–315.

Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H. & DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology*, **178**, 591–599.

Sullivan, C. W. & Palmisano, A. C. (1984). Sea ice microbial communities: distribution, abundance, and diversity of ice bacteria in McMurdo Sound, Antarctica, in 1980. *Applied and Environmental Microbiology*, **47**, 788–795.

Svartström, O., Alneberg, J., Terrapon, N., Lombard, V., de Bruijn, I., Malmsten, J., Dalin, A.-M., El Muller, E., Shah, P., Wilmes, P., Henrissat, B., Aspeborg, H. & Andersson, A. F. (2017). Ninety-nine de novo assembled genomes from the moose (*Alces alces*) rumen microbiome provide new insights into microbial plant biomass degradation. *The ISME Journal*, **11**, 2538–2551.

Swan, B. K., Tupper, B., Sczyrba, A., Lauro, F. M., Martinez-Garcia, M. et al. (2013). Prevalent genome streamlining and latitudinal divergence of planktonic bacteria in the surface ocean. *Proceedings of the National Academy of Sciences of the United States of America*, **110**, 11463–11468.

Sørensen, I., Pettolino, F. A., Bacic, A., Ralph, J., Lu, F., O'Neill, M. A., Fei, Z., Rose, J., Domozych, D. S. & Willats, W. G. (2011). The charophycean green algae provide insights into the early origins of plant cell walls. *The Plant Journal*, **68**, 201–211.

Tada, Y., Taniguchi, A., Nagao, I., Miki, T., Uematsu, M., Tsuda, A. & Hamasaki, K. (2011). Differing growth responses of major phylogenetic groups of marine bacteria to natural phytoplankton blooms in the western North Pacific Ocean. *Applied and Environmental Microbiology*, **77**, 4055–4065.

Tamura, K., Hemsworth, G. R., Déjean, G., Rogers, T. E., Pudlo, N. A., Urs, K., Jain, N., Davies, G. J., Martens, E. C. & Brumer, H. (2017). Molecular mechanism by which prominent human gut *Bacteroidetes* utilize mixed-linkage beta-glucans, major health-promoting cereal polysaccharides. *Cell Reports*, **21**, 417–430.

Tan, S., Zhou, J., Zhu, X., Yu, S., Zhan, W., Wang, B. & Cai, Z. (2015). An association network analysis among microeukaryotes and bacterioplankton reveals algal bloom dynamics. *Journal of Phycology*, **51**, 120–132.

Tang, K., Jiao, N., Liu, K., Zhang, Y. & Li, S. (2012). Distribution and functions of TonB-dependent transporters in marine bacteria and environments: implications for dissolved organic matter utilization. *PLoS ONE*, **7**, e41204.

Tang, K., Lin, Y., Han, Y. & Jiao, N. (2017). Characterization of potential polysaccharide utilization systems in the marine *Bacteroidetes Gramella flava* JLT2011 using a multi-omics approach. *Frontiers in Microbiology*, **8**, 220.

Taylor, J. D., Cottingham, S. D., Billinge, J. & Cunliffe, M. (2014). Seasonal microbial community dynamics correlate with phytoplankton-derived polysaccharides in surface coastal waters. *The ISME Journal*, **8**, 245–248.

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A. et al. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science*, **336**, 608–611.

Teeling, H., Fuchs, B. M., Bennke, C. M., Krüger, K., Chafee, M., Kappelmann, L., Reintjes, G., Waldmann, J., Quast, C., Glöckner, F. O., Lucas, J., Wichels, A., Gerdts, G., Wiltshire, K. H. & Amann, R. I. (2016). Recurring patterns in bacterioplankton dynamics during coastal spring algae blooms. *eLife*, **5**, e11888.

Teeling, H., Meyerdierks, A., Bauer, M., Amann, R. & Glöckner, F. O. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environmental Microbiology*, **6**, 938–947.

Terrapon, N., Lombard, V., Drula, E., Coutinho, P. & Henrissat, B. (2017). The CAZy database/the Carbohydrate-Active Enzyme (CAZy) database: principles and usage guidelines. In K. Aoki-Kinoshita, ed., *A Practical Guide to Using Glycomics Databases*, 117–131. Springer.

Terrapon, N., Lombard, V., Drula, E., Lapébie, P., Al-Masaudi, S., Gilbert, H. J. & Henrissat, B. (2018). PULDB: the expanded database of Polysaccharide Utilization Loci. *Nucleic Acids Research*, **46**, D677–D683.

Terrapon, N., Lombard, V., Gilbert, H. J. & Henrissat, B. (2015). Automatic prediction of polysaccharide utilization loci in *Bacteroidetes* species. *Bioinformatics*, **31**, 647–655.

The UniProt Consortium. (2014). Activities at the Universal Protein Resource (UniProt). *Nucleic Acids Research*, **42**, D191–D198.

Thiele, S., Fuchs, B. M. & Amann, R. I. (2011). Identification of microorganisms using the ribosomal RNA approach and fluorescence in situ hybridization. In P. Wilderer, ed., *Treatise on Water Science, Vol 3: Aquatic Chemistry and Biology*, 171–189. Elsevier Science.

Thiele, S., Fuchs, B. M., Ramaiah, N. & Amann, R. (2012). Microbial community response during the iron fertilization experiment LOHAFEX. *Applied and Environmental Microbiology*, **78**, 8803–8812.

Thomas, F., Barbeyron, T., Tonon, T., Génicot, S., Czjzek, M. & Michel, G. (2012). 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*. *Environmental Microbiology*, **14**, 2379–2394.

Thomas, F., Hehemann, J.-H., Rebuffet, E., Czjzek, M. & Michel, G. (2011). Environmental and gut *Bacteroidetes*: the food connection. *Frontiers in Microbiology*, **2**, 93.

Tonon, T., Li, Y. & McQueen-Mason, S. (2017). Mannitol biosynthesis in algae: more widespread and diverse than previously thought. *New Phytologist*, **213**, 1573–1579.

Tuncil, Y. E., Xiao, Y., Porter, N. T., Reuhs, B. L., Martens, E. C., Hamaker, B. R., Walter, J. & Ruby, E. G. (2017). Reciprocal prioritization to dietary glycans by gut bacteria in a competitive environment promotes stable coexistence. *mBio*, **8**, e01068–17.

Turley, C. M. & Stutt, E. D. (2000). Depth-related cell-specific bacterial leucine incorporation rates on particles and its biogeochemical significance in the Northwest Mediterranean. *Limnology and Oceanography*, **45**, 419–425.

Turner, J. T. (2015). Zooplankton fecal pellets, marine snow, phytodetritus and the ocean's biological pump. *Progress in Oceanography*, **130**, 205–248.

Tuson, H. H., Foley, M. H., Koropatkin, N. M. & Biteen, J. S. (2018). The starch utilization system assembles around stationary starch-binding proteins. *Biophysical Journal*, **114**, 242–250.

Tyanova, S., Temu, T., Sinitcyn, P., Carlson, A., Hein, M. Y., Geiger, T., Mann, M. & Cox, J. (2016). The Perseus computational platform for comprehensive analysis of (prote)omics data. *Nature Methods*, **13**, 731–740.

Ulaganathan, T., Shi, R., Yao, D., Gu, R. X., Garron, M. L., Cherney, M., Tieleman, D. P., Sterner, E., Li, G., Li, L., Linhardt, R. J. & Cygler, M. (2017). Conformational flexibility of PL12 family heparinases: structure and substrate specificity of heparinase III from *Bacteroides thetaiotaomicron* (BT4657). *Glycobiology*, **27**, 176–187.

Unfried, F., Becker, S., Robb, C. S., Hehemann, J. H., Markert, S. et al. (2018). Adaptive mechanisms that provide competitive advantages to marine bacteroidetes during microalgal blooms. *The ISME Journal*, **12**, 2894–2906.

Urbani, R., Magaletti, E., Sist, P. & Cicero, A. M. (2005). Extracellular carbohydrates released by the marine diatoms *Cylindrotheca closterium*, *Thalassiosira pseudonana* and *Skeletonema costatum*: effect of P-depletion and growth status. *Science of the Total Environment*, **353**, 300–306.

Valdehuesa, K. N. G., Ramos, K. R. M., Moron, L. S., Lee, I., Nisola, G. M., Lee, W.-k. & Chung, W.-j. (2018). Draft genome sequence of newly isolated agarolytic bacteria *Cellulophaga omnivescoria* sp. nov. W5C carrying several gene loci for marine polysaccharide degradation. *Current Microbiology*, **75**, 925–933.

Van Geel-Schutten, G. H., Faber, E. J., Smit, E., Bonting, K., Smith, M. R., Ten Brink, B., Kamerling, J. P., Vliegenthart, J. F. G. & Dijkhuizen, L. (1999). Biochemical and structural characterization of the glucan and fructan exopolysaccharides synthesized by the *Lactobacillus reuteri* wild-type strain and by mutant strains. *Applied and Environmental Microbiology*, **65**, 3008–3014.

Vincent, A. T., Derome, N., Boyle, B., Culley, A. I. & Charette, S. J. (2017). Next-generation sequencing (NGS) in the microbiological world: how to make the most of your money. *Journal of Microbiological Methods*, **138**, 60–71.

Vizcaíno, J. A., Csordas, A., Del-Toro, N., Dianes, J. A., Griss, J., Lavidas, I., Mayer, G., Perez-Riverol, Y., Reisinger, F., Ternent, T., Xu, Q. W., Wang, R. & Hermjakob, H. (2016). 2016 update of the PRIDE database and its related tools. *Nucleic Acids Research*, **44**, 11033.

Voget, S., Wemheuer, B., Brinkhoff, T., Vollmers, J., Dietrich, S., Giebel, H.-A., Beardsley, C., Sardemann, C., Bakenhus, I., Billerbeck, S., Daniel, R. & Simon, M. (2015). Adaptation of an abundant *Roseobacter* RCA organism to pelagic systems revealed by genomic and transcriptomic analyses. *The ISME Journal*, **9**, 371–384.

Vollmers, J., Wiegand, S. & Kaster, A. K. (2017). Comparing and evaluating metagenome assembly tools from a microbiologist's perspective - not only size matters! *PLoS ONE*, **12**, e0169662.

Weiss, M., Abele, U., Weckesser, J., Welte, W., Schiltz, E. & Schulz, G. (1991). Molecular architecture and electrostatic properties of a bacterial porin. *Science*, **254**, 1627–1630.

Wemheuer, B., Wemheuer, F., Hollensteiner, J., Meyer, F.-D., Voget, S. & Daniel, R. (2015). The green impact: bacterioplankton response toward a phytoplankton spring bloom in the southern North Sea assessed by comparative metagenomic and metatranscriptomic approaches. *Frontiers in Microbiology*, **6**, 805.

Wilhelm, L. J., Tripp, H. J., Givan, S. A., Smith, D. P. & Giovannoni, S. J. (2007). Natural variation in SARII marine bacterioplankton genomes inferred from metagenomic data. *Biology Direct*, **2**, 27.

Williams, T. J., Wilkins, D., Long, E., Evans, F., DeMaere, M. Z., Raftery, M. J. & Cavicchioli, R. (2013). The role of planktonic *Flavobacteria* in processing algal organic matter in coastal East Antarctica revealed using metagenomics and metaproteomics. *Environmental Microbiology*, **15**, 1302–1317.

Wiltshire, K. H., Kraberg, A., Bartsch, I., Boersma, M., Franke, H.-D., Freund, J., Gebuehr, C., Gerdts, G., Stockmann, K. & Wichels, A. (2010). Helgoland Roads, North Sea: 45 years of change. *Estuaries and Coasts*, **33**, 295–310.

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

Woyke, T., Xie, G., Copeland, A., González, J. M., Han, C., Kiss, H., Saw, J. H., Senin, P., Yang, C., Chatterji, S., Cheng, J.-F., Eisen, J. A., Sieracki, M. E. & Stepanauskas, R. (2009). Assembling the marine metagenome, one cell at a time. *PLoS ONE*, **4**, e5299.

Wu, J., Lv, Y., Liu, X., Zhao, X., Jiao, G., Tai, W., Wang, P., Zhao, X., Cai, C. & Yu, G. (2015). Structural study of sulfated fuco-oligosaccharide branched glucuronomannan from *Kjellmaniella crassifolia* by ESI-CID-MS/MS. *Journal of Carbohydrate Chemistry*, **34**, 303–317.

Wu, Y.-W., Tang, Y.-H., Tringe, S. G., Simmons, B. A. & Singer, S. W. (2014). MaxBin: an automated binning method to recover individual genomes from metagenomes using an expectation-maximization algorithm. *Microbiome*, **2**, 26.

Wustman, B. A., Lind, J., Wetherbee, R. & Gretz, M. R. (1998). Extracellular matrix assembly in diatoms (*Bacillariophyceae*) - III. Organization of fucoglucurono-galactans within the adhesive stalks of *Achnanthes longipes*. *Plant Physiology*, **116**, 1431–1441.

Xing, P., Hahnke, R. L., Unfried, F., Markert, S., Huang, S., Barbeyron, T., Harder, J., Becher, D., Schweder, T., Glöckner, F. O., Amann, R. I. & Teeling, H. (2015). Niches of two polysaccharide-degrading *Polaribacter* isolates from the North Sea during a spring diatom bloom. *The ISME Journal*, **9**, 1410–1422.

Yamaguchi, T., Kawakami, S., Hatamoto, M., Imachi, H., Takahashi, M., Araki, N., Yamaguchi, T. & Kubota, K. (2015). In situ DNA-hybridization chain reaction (HCR): a facilitated in situ HCR system for the detection of environmental microorganisms. *Environmental Microbiology*, **17**, 2532–2541.

Yang, C., Li, Y., Zhou, B., Zhou, Y., Zheng, W., Tian, Y., Van Nostrand, J. D., Wu, L., He, Z., Zhou, J. & Zheng, T. (2015). Illumina sequencing-based

analysis of free-living bacterial community dynamics during an *Akashiwo sanguine* bloom in Xiamen sea, China. *Scientific Reports*, **5**, 8476.

Yarza, P., Yilmaz, P., Pruesse, E., Glöckner, F. O., Ludwig, W., Schleifer, K.-H., Whitman, W. B., Euzéby, J., Amann, R. & Rosselló-Móra, R. (2014). Uniting the classification of cultured and uncultured bacteria and archaea using 16S rRNA gene sequences. *Nature Reviews Microbiology*, **12**, 635–645.

Yilmaz, P., Kottmann, R., Field, D., Knight, R., Cole, J. R. et al. (2011). Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nature Biotechnology*, **29**, 415–420.

Yin, Y., Mao, X., Yang, J., Chen, X., Mao, F. & Xu, Y. (2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. *Nucleic Acids Research*, **40**, W445–W451.

Yool, A. & Tyrrell, T. (2003). Role of diatoms in regulating the ocean's silicon cycle. *Global Biogeochemical Cycles*, **17**, https://doi.org/10.1029/2002GB002018.

Yu, C.-S., Chen, Y.-C., Lu, C.-H. & Hwang, J.-K. (2006). Prediction of protein subcellular localization. *Proteins - Structure Function and Bioinformatics*, **64**, 643–651.

Yu, N. Y., Wagner, J. R., Laird, M. R., Melli, G., Rey, S., Lo, R., Dao, P., Sahinalp, S. C., Ester, M., Foster, L. J. & Brinkman, F. S. L. (2010). PSORTb 3.0: improved protein subcellular localization prediction with refined localization subcategories and predictive capabilities for all prokaryotes. *Bioinformatics*, **26**, 1608–1615.

Zhang, Y., Wen, Z., Washburn, M. P. & Florens, L. (2010). Refinements to label free proteome quantitation: how to deal with peptides shared by multiple proteins. *Analytical Chemistry*, **82**, 2272–2281.

Zhang, Z., Chen, Y., Wang, R., Cai, R., Fu, Y. & Jiao, N. (2015). The fate of marine bacterial exopolysaccharide in natural marine microbial communities. *PLoS ONE*, **10**, e0142690.

Zhou, J., Bruns, M. A. & Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *Applied and Environmental Microbiology*, **62**, 316–322.

Zhu, Y., Suits, M. D. L., Thompson, A. J., Chavan, S., Dinev, Z., Dumon, C., Smith, N., Moremen, K. W., Xiang, Y., Siriwardena, A., Williams,

S. J., Gilbert, H. J. & Davies, G. J. (2010). Mechanistic insights into a Ca^{2+}-dependent family of alpha-mannosidases in a human gut symbiont. *Nature Chemical Biology*, **6**, 125–132.

Zubkov, M. V., Fuchs, B. M., Tarran, G. A., Burkill, P. H. & Amann, R. (2003). High rate of uptake of organic nitrogen compounds by *Prochlorococcus* cyanobacteria as a key to their dominance in oligotrophic oceanic waters. *Applied and Environmental Microbiology*, **69**, 1299–1304.

Acknowledgements

First of all, I would like to thank my supervisors Dr. Hanno Teeling and Prof. Dr. Rudolf Amann for giving me the opportunity to work in the Molecular Ecology Group. I am very grateful that I was able to work on such a big, but sometimes also overwhelming dataset. Thank you for your guidance and for giving me the time to develop both my ecological thinking and my bioinformatic skills.

Thank you Prof. Dr. Carol Arnosti for reviewing my thesis, for fruitful discussions at meetings and of course for travelling to Bremen for my PhD defense.

I would also like to thank Prof. Dr. Michael Friedrich and Dr. Jan-Hendrik Hehemann for accepting to be part of my examination board.

Thank you Anissa and Taylor for taking part in my examination board as student members.

Further, I would like to thank everyone involved in my scientific development at MPI. Thank you to everyone involved in my thesis committee meetings: Prof. Dr. Rudolf Amann, Dr. Hanno Teeling, PD Dr. Bernhard Fuchs, Dr. Harald Gruber-Vodicka, Dr. Jan-Hendrik Hehemann and Dr. Luis Humberto Orellana Retamal. And to everyone involved in the COGITO and POMPU projects. Thank you for the fruitful discussions and the scientific advice.

I would also like to thank everyone from the Molecular Ecology Group for always being helpful and providing a nice working atmosphere. I very much enjoyed our group retreats and regular cake meetings.

Thank you to all previous and present members of the Teeling Group. I would especially like to thank Meghan and Ben for sharing the struggle we had with these 38 metagenomes and for your moral and scientific support.

Further, I would also like to thank everyone involved in thesis proof-reading.

Additionally, I would like to thank the MarMic program and everyone involved. Thank you Dr. Christiane Glöckner for taking care of all MarMic organizational matters.

And finally, I would like to thank my family and friends. Your love, friendship and support have helped me to get through this. Thank you Joakim, Ulrike, Knut, Julia, Anissa, Laura, the Back- and Spielequeens and everybody else. Dankeschön! Tack så mycket!

"and now we get to the hard part, the endings, the farewells, and the famous last words. If you don't hear from me often, remember that you are in my thoughts." – Paul Auster, Moon Palace

Universität Bremen

Name: Karen Krüger

Anschrift: Georg-Gröning-Straße 34, 28209 Bremen

Ort, Datum: Bremen, 30.01.2019

ERKLÄRUNG

Hiermit erkläre ich, dass ich die Doktorarbeit mit dem Titel:

Polysaccharide utilization loci and associated genes in marine *Bacteroidetes* – compositional diversity and ecological relevance

selbstständig verfasst und geschrieben habe und außer den angegebenen Quellen keine weiteren Hilfsmittel verwendet habe.

Ebenfalls erkläre ich hiermit, dass es sich bei den von mir abgegebenen Arbeiten um drei identische Exemplare handelt.

_____

(Unterschrift)