[illegible]

Dissertation

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

Dem Fachbereich Biologie/Chemie der
Universit□t Bremen
vorgelegt von

Sixing Huang

Bremen
M□rz 2013

□

Die vorliegende Arbeit wurde in der Zeit von August 2009 bis M□rz 2013 am Max-Planck-Institut f□r Marine Mikrobiologie in Bremen angefertigt.

1. Gutachter: Prof. Dr. Rudolf Amann
2. Gutachter: Prof. Dr. Karl-Heinz Blotevogel
3. Pr□fer: Prof. Dr. Ulrich Fischer
4. Pr□fer: Dr. Hanno Teeling

Tag des Promotionskolloquiums: 23.04.2013

S□□ □ a□□ 3

Z□□a□ □ e□fa□□□□g 4

Li□□□f abb□e□ia□i□□□ 5

1. I□□□□d□c□i□□ 6
1.1 The marine food web 6
1.2 Carbohydrates in marine ecosystems 8
1.3 Carbohydrate-active enzymes (CAZymes) 10
1.4 Enzymatic synergisms of CAZymes 12
1.5 Trident: a bioinformatic pipeline for automatic CAZyme annotation 14
1.6 CAZyme profiling can reveal an organism's lifestyle 16
1.7 Combined taxonomic classification and CAZyme profiling can reveal trophic relations in metagenomes 18
1.7.1 Taxonomic classification of metagenomes 18
1.7.2 CAZyme analysis of taxonomically classified sequences 21
1.8 Metagenomes from the Microbial Interactions in Marine Systems (MIMAS) Project 21
1.9 Metagenomes from the Logatchev deep-sea hydrothermal vent field 23
1.10 The aims of this thesis: establishment and application of CAZyme profiling for studies in microbial ecology 24

2. P□b□ica□i□□□ a□d Ma□□□c□i□□□ 26
2.1 Practical application of Self-Organizing Maps to interrelate biodiversity and functional data in NGS-based metagenomics 27
2.2 Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom 39
2.3 Carbohydrate-active enzyme profiling of a diatom-induced bacterioplankton succession reveals niche adaptations and trophic connections 44
2.4 Metagenomics reveals niche-differentiation and habitat-specific adaptation in surface sediments of the Logatchev hydrothermal field 88
2.5 *H□□□□□□□□□□ □□□□ □□□□*: Complete genome sequencing and proteomics identify the first cultivated euryarchaeon from a deep-sea anoxic brine lake as polysaccharide degrader 135

3. Di□c□□□i□□ 190
3.1 The accuracy of CAZyme identification 190
3.2 The necessity of characterizing carbohydrate metabolism genes other than CAZymes 192
3.3 CAZymes provided insight into the carbohydrate flow in the bacterioplankton bloom in 2009 193
3.4 The different CAZyme profiles in *F□□□□□□□□□□□□□□□□* reflected their niche adaptations 199
3.5 The CAZyme profiles from the Logatchev metagenome showed that □-glucans and glycosylation are common among the deep-sea microbes. 204
3.6 CAZyme profiling as tool for the study of microbial interactions in meta-omics 206

4. O□□□□□□ 207

5. Refe□e□ce□ ci□ed i□ I□□□□d□c□i□□, Di□c□□□i□□ a□d O□□□□□□ 210

6. Ac□□□□ □edge□ e□□□ 225

7. A□□e□di□ 227
7.1 The gill chamber epibiosis of deep-sea □□□ □□□□□ □□□□□□□□□ shrimp thoroughly investigated by metagenomics and discovery of zetaproteobacterial epibionts 227
7.2 Complete genome sequence of the algae-associated marine flavobacterium *F□□□ □□□ □□□□□□□□□□* KMM 3901^T 229
7.3 Geomicrobiology of Hot Lake, a shallow-sea hydrothermal vent site off Panarea Island, Italy 231

S□□ □ a□□

Carbohydrate-active enzymes (CAZymes) account on average for only 1-2% of the bacterial genes [1-4], yet they are responsible for the metabolism of carbohydrates (sugars) - one of the most important class of biological macromolecules. Besides their functions as structure and storage compounds, sugar molecules act in various other functions, such as protein stabilization and osmoregulation. Sugars also constitute important intermediates in the food web, can function as signal molecules and serve as precursors for biosyntheses. It is becoming more and more common to address the carbohydrate turnover in modern environmental microbiology studies. However, it is hard to directly assess the sugar composition and concentrations on a cellular level, let alone their □□□□ turnover rates. Despite these technical difficulties, it is still possible to characterize the carbohydrate metabolisms indirectly □ via the study of their metabolic genes, namely CAZymes. This is possible because CAZymes catalyze the turnover of sugars. To be specific, they are synthesized by glycosyltransferases (GT) and degraded by glycoside hydrolases (GH), polysaccharide lyases (PL) and carbohydrate esterases (CE). Therefore, gene frequencies and expression levels of CAZymes should be indicators of the □□□□ availability of distinct sugars.

The current knowledge on CAZymes is largely derived from carbohydrate synthesis and degradation of terrestrial plants, whereas their marine counterparts are still largely unknown. This asymmetry is also referred to as the "knowledge gap" of marine CAZymes. This limits not only our interpretation of the 'omics (in particular metagenomic) data and understanding of marine ecosystems, but it also keeps us from utilizing this vast enzyme repertoire for biotechnological purposes. This doctoral project focuses on CAZyme distributions in marine genomic and metagenomic studies with a focus on the MIMAS (Microbial Interactions in Marine Systems) Project and a sediment sample from the Logatchev hydrothermal vent site. Although each project has been described in previous studies, the carbohydrate metabolisms in these habitats were still largely under-researched. At first, the metagenome sequences were taxonomically classified and clustered into 'taxobins'. Afterwards, I studied the carbohydrate metabolic capacities of the dominant taxobins. In the MIMAS study, I also discuss possible trophic relations among different taxa based on their CAZymes profiles and extend the discussion to the niche adaptation of *F*□□□□□□□□□□□□□□□□. In summary, this thesis demonstrates the usefulness of CAZyme profiling as a tool for interpreting 'omics data in microbial ecology. This thesis constitutes the first attempt so far to apply CAZyme analyses to elucidate a multi-level food web as well as to characterize the carbohydrate metabolism in a deep-sea habitat. Such studies are necessary for an in-depth understanding of the marine carbon cycle and could also provide a guideline for the selection of promising candidate CAZymes for industrial applications.

Z□□a□ □e□fa□□□□g

Im Durchschnitt sind 1-2% der Gene eines bakteriellen Genoms sog. □□□□□□□□□□□□-□□□□□ □□□□□ □□ (CAZymes) [1-4]. Dies sind Gene, die f□r den Polysaccharid-Stoffwechsel verantwortlich sind. Polysaccharide sind wichtige biologische Makromolek□le. Sie dienen nicht nur als Struktur- und Speicherstoffe, sondern tragen auch zur Stabilisierung und zur Osmoregulation bei. Oligosaccharide stellen dar□ber hinaus wichtige Intermediate in der Nahrungskette dar, dienen als Signalmolek□le und sind Ausgangsstoffe f□r zahlreiche Biosynthesen. Untersuchungen von Polysacchariden sind daher immer h□ufiger Bestandteil umweltmikrobiologischer Studien. Es f□llt jedoch naturgem□□ schwer, die Zusammensetzung oder die Konzentration eines Polysaccharids oder Zuckermonomers auf zellul□rem Niveau zu messen, ganz zu schweigen von seiner Umwandlungsrate. Trotzdem ist eine indirekte Untersuchung via CAZymes m□glich, weil CAZymes die Auf-, Ab- und Umbaureaktionen von Polysacchariden katalysieren. Polysaccharide werden von Glycosyltransferasen (GT) synthetisiert und von Glycosid-Hydrolasen (GH), Polysaccharid-Lyasen (PL) und Kohlenhydrat-Esterasen (carbohydrate esterases, CE) abgebaut. Die Genh□ufigkeiten und Expressionsniveaus von CAZymes lassen daher indirekte R□ckschl□sse auf die von ihnen metabolisierten Polysaccharide zu.

Der □berwiegende Teil der Information □ber CAZymes beruht auf Untersuchungen an Landpflanzen. □ber marine Polysaccharide, die in gro□en Mengen von Algen produziert werden, ist hingegen nur wenig bekannt. Diese Wissensl□cke stellt eine Einschr□nkung bei der Interpretation von 'Omics-Daten dar, wie sie insbesondere die moderne Metagenomik erzeugt und limitiert unser Verst□ndnis von marinen □kosystemen. Dies gilt in gleichem Ma□e f□r m□gliche biotechnologische Anwendungen mariner CAZymes. Diese Doktorarbeit konzentriert sich daher auf CAZyme-Analysen, insbesondere von Metagenomen, die im Rahmen zweier Projekte gewonnen wurden, n□mlich des MIMAS-Projekts (□ □□□□□□□□□□/□□□□□□□□□□□□□ □□ □ □□□□□□ □□□□□□□ □) sowie einer Studie einer Sedimentprobe aus dem Logatchev Hydrothermal Feld. Die metagenomischen Sequenzen wurden zuerst taxonomisch klassifiziert und dadurch in sog. □Taxobins□ geclustert. Dabei fokussiert sich diese Doktorarbeit nicht nur auf die Stoffwechselwege innerhalb der Taxobins, sondern auch darauf, ob sich auf diese Weise trophische Beziehungen zwischen Taxa ableiten lassen. Eine wesentlich Motivation war also, die Tauglichkeit von CAZyme-Profiling als Werkzeug in 'omics-Studien n□her zu untersuchen. In diesem Zusammenhang wurde das erste Mal versucht, mit Hilfe von CAZyme-Analysen eine Nahrungskette □ber mehrere Niveaus nachzuverfolgen und Nischenanpassung bakterioplanktischer *F*□□□□□□□□□□□□□□□ zu verstehen. Dar□ber hinaus enth□lt diese Arbeit die erste Studie von CAZymes in einem Tiefsee-Habitat. Solche Studien sind f□r ein vertieftes Verst□ndnis des marinen Kohlenstoffkreislaufs n□tig und k□nnten zudem zur sinnvollen Selektion biotechnologisch n□tzlicher hydrolytischer Enzyme beitragen.

List of abbreviations

CARD-FISH	Catalyzed Reporter Deposition Fluorescence In-Situ Hybridization
CAZy	Carbohydrate-active enzymes (database)
CAZymes	Carbohydrate-active enzymes
CE	Carbohydrate esterase
DNA	Deoxyribonucleic acid
DOM	Dissolved organic matter
e.g.	exempli gratia
FISH	Fluorescence In-Situ Hybridization
Flavo. group A	*Flavo.* sp. Hel3_A1_48
Flavo. group B	*Flavo.* sp. Hel1_33_131
GH	Glycoside hydrolase(s)
GT	Glycosyltransferase(s)
LHF	Logatchev hydrothermal vent field
MIMAS	Microbial Interactions in Marine Systems
NPP	Net primary production
OM	Organic matter
ORF	Open reading frame
PL	Polysaccharide lyase
Polaribacter sp. 49	Polaribacter sp. Hel1_33_49
Polaribacter sp. 85	Polaribacter sp. Hel1_85
POM	Particulate organic matter
Psychroserpens sp. D35	Psychroserpens sp. Hel1_31_5_D35
TEP	Transparent exopolymers

1. Introduction

1.1 The marine food web

Oceans cover 71% of the Earth's surface and contain 97% of its water [5]. These gigantic water bodies not only have an enormous influence on Earth's climate, but function also as an ecosystem. According to Mora et al., there are about 8.7 million predicted species on Earth and about one fourth of them live in the oceans. More than 90% of these marine species have yet to be described [6]. Marine ecosystems rely on the sun's energy or chemical energy to support the growth and metabolism of its inhabitants. Autotrophic organisms use this energy to convert inorganic materials into organic molecules such as carbohydrates. The annual global primary production has been estimated to be 105 Pg (105×10^{15} g) [7], about 50% of which are attributed to marine algae [8]. In the oceans, diatoms are among the most important primary producers. Diatoms account for less than one third of primary production in oligotrophic areas but they predominate in the highly productive oceanic regions [7]. Together, diatoms are estimated to fix 20 Pg carbon annually [7, 9, 10], which corresponds to about 20-23% of the global NPP [7-9, 11]. Diatoms' appearances are seasonal. This culminates in annually recurring, sometimes massive diatom blooms that characterize the upwelling zones and continental shelves in higher latitudes worldwide [12-15]. In the deep-sea, various chemosynthetic bacteria such as sulfur-oxidizing *Gamma-* and *Epsilonproteobacteria* form the basis of the food web [16-18].

These tiny primary producers support a myriad of heterotrophic organisms as small as unicellular flagellates and ciliates and as large as fish and marine mammals in the oceans. They serve as food source for various zooplankton and the latter in turn fall prey to bigger predators. Debris, secreted materials and fecal pellets become nutrient sources for decomposing organisms. In addition, viral lysis releases cellular materials back to the environment [19, 20]. The organic matter (OM) released from the loop is either insoluble or soluble. The insoluble portion becomes particulate organic matter (POM). POM generally has a size larger than 1 µm [21]. From the photic upper layer of the water column, POM sinks continuously to deeper aphotic layers. This phenomenon is reminiscent of snowfall. For this reason, sinking particles visible to the human eye are referred to as "marine snow". Marine snow is one of the most important nutrient sources for the organisms in the aphotic zone. In contrast, the soluble portion of the

material is called DOM. The sizes of DOM are in micro- and nanometer scales. Examples of DOM are bacterial glycogens and algal laminarins.

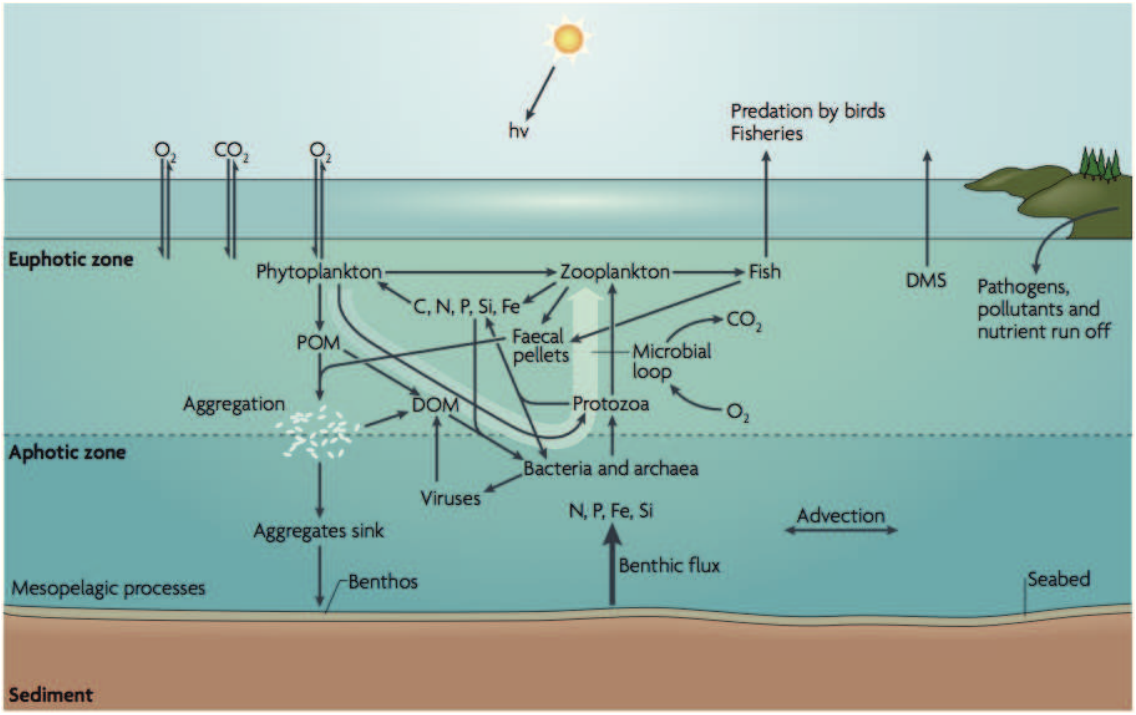


Figure 1. The recycling of inorganic and organic materials in marine ecosystems [21]. DMS, dimethylsulfide; hv, light; POM, particulate organic matter; DOM, dissolved organic matter.

Bacteria dominate the recycling of POM and DOM and interact with the marine ecosystems in multiple ways [21]. They not only attach to the POM surface and degrade POM, but also are capable of solubilizing POM enzymatically to DOM [22], which is accessible to microbes [21]. About 10% of all the DOM could be assembled as gels, which are three-dimensional networks of biopolymers imbedded in seawater [23, 24]. The algal extracellular and cell wall matrices are examples of gels [25]. Microorganisms are able to attach to the surfaces of gel-form DOM and degrade the matters with extracellular hydrolases [21]. Thus, bacteria are in a central position of the marine food web and the microbial loop. On the one hand, they are decomposers that disintegrate complex materials from other trophic levels and recycle the degradation products back to the environment. On the other hand, they fall prey to protozoa. Nutrients then go through levels of zooplankton and end up in fish. This is called the microbial loop [20, 21]. About 90-95% of the DOM is recalcitrant to bacterial degradation [21] (Figure 1).

1.2 Carbohydrates in marine ecosystems

In primary production, energy-rich organic compounds such as sugars are synthesized. Carbohydrates, also known as saccharides or sugars, are an essential building block of life. Carbohydrates consist mostly of carbon, hydrogen and oxygen in an approximate ratio of CH_2O. Sugar molecules can also contain heteroatoms such as nitrogen, phosphorus and sulfur. Chemically, monosaccharides are polyhydroxy aldehydes and ketones. Two monosaccharides can form a disaccharide by condensation. The resulting chemical bond is termed glycosidic bond. There are two types of glycosidic bonds, designated as □ and □. They represent the two possible configurations between the anomeric center and the anomeric reference atom of the first monomer. The monomer is □ and the bond is also □ if the two oxygen atoms are cis. The monomer and the bond are □ if the two oxygen atoms are trans. For example, a maltose molecule is formed by two glucose monomers via an □-1,4-glycosidic bond. The numbering □1,4□indicates that the bond is located at the C1 atom of the first monomer and the C4 atom of the second monomer [26].

Multiple monomers can undergo the same type of reaction to form oligosaccharides and polysaccharides. These carbohydrate polymers are sometimes named based on their types of monomers and glycosidic bonds. In polysaccharides, the glycosidic bonds are formed at the anomeric center of the first monomer and at any non-anomeric carbon atom of the second monomer. One example is □-glucan, which consists of multiple D-glucose monomers that are linked via □-glycosidic bonds. The possible permutations of oligosaccharides are astronomical. First, the monosaccharide alphabet is extensive. A D-aldohexose alone has three chiral centers that allow up to 2^3 = 8 stereoisomers (allose, altrose, glucose, mannose, gulose, idose, galactose and talose) and this value is even larger when the derivatives are taken into account. Second, the □ and □ configurations double the possible combinations. Third, glycosidic bonds can be formed on different carbon atoms. For instance, except the C5, all the other five carbon atoms in a pyranose can form glycosidic bonds. Last but not least, polysaccharides can contain branches [26]. The variability of glycosidic bonds lead to the vast stereochemical variations in carbohydrates. Theoretically, a hexasaccharide can have up to 10^{12} possible linear or branched isomers [27]. For this reason, polysaccharides can store information several orders of magnitude higher than any other biological molecules of the same length (For comparison, a hexaoligopeptide can have 20^6 □ 6 □ 10^7 possible permutations, while a hexanucleotide has only 46 = 4096 permutations.) [27]. In nature, however, only a subset of these permutations does occur. For example, among the eight stereoisomers of D-aldohexose, idose is not

found in nature while allose, altrose, gulose and talose are very rare. In addition, only certain kinds of glycosidic bonds are formed in some well-characterized polysaccharides. Also, there are regular branching patterns. For example, glycogens are □-1,4- and □-1,6-linked glucoses. The glycogen molecule starts a branch about every ten glucoses with an □-1,6 linkage [26].

It is technologically difficult to characterize carbohydrates with these huge amounts of isomers. This phenomenon is called □isomer barrier□ [27]. The research in carbohydrates is still in its early stage. It is heavily driven by human interest and the effort is not balanced. In contrast to their well-studied terrestrial counterparts, marine polysaccharides are still under-researched with respect to their compositions, structures and functions. Characteristic marine polysaccharides are known to exist in marine algae in the form of cell wall matrix components and storage compounds. Some of these matrix polysaccharides in algae are highly anionic due to sulfate or carboxylate groups. Examples for sulfated polysaccharides are fucans [28], carrageenans [29], agars [30, 31], porphyrans [31], naviculan [32] and ulvans [33], while alginate is a carboxylated polysaccharide [34]. In contrast, the anionic polysaccharide is largely absent in land and freshwater plants, which probably have lost this feature as they adapted from the highly sulfated seawater (28 mM) to the low-sulfate freshwater (ranges from 0.09 to 1.4 mM) [34]. Anionic polysaccharides have multiple functions. It has been suggested that the negative charges in anionic marine polysaccharides play a role in salt-resistance [35, 36]. Furthermore, sulfated polysaccharides are more resistant to enzymatic breakdown because the extra sulfate groups have to be removed by sulfatases [37]. In diatoms, sulfated polysaccharides are found in the form of transparent exopolymer particles (TEP). They are POM deposited outside of the diatom cells that lead to diatom flocculation. The latter accelerates diatom sinking and thus removes diatoms from the productive surface water layer. Therefore, TEP is considered to be one of termination factors of diatom blooms [38].

Compared to plants, which synthesize sucrose and starch as their primary photosynthetic polysaccharides, diatoms use chrysolaminarins [39]. Chrysolaminarin generally constitute 10 to 50% of the cellular carbon in the exponentially growing diatoms [10, 40, 41] and up to 70% in *C*□□□□□□□□□□ □□□□□□□□□□□□□□□ [42]. Based on this knowledge and given that diatoms are responsible for about one fourth of the global primary production [7, 8], the annual chrysolaminarin turnover is estimated to be in the order of several petagrams (or gigatons) [7]. Although chrysolaminarins resemble laminarins structurally, they do not have mannitol at their termini. In fact, no mannitol pathway could be detected in diatoms so far [43]. Chrysolaminarins are

composed of □-1,3-linked linear sections and □-1,6-linked branches [39]. These branching structures are reminiscent of the □-linked starch and glycogen, which are highly branched, compact and serve as storage compounds in plants and animals.

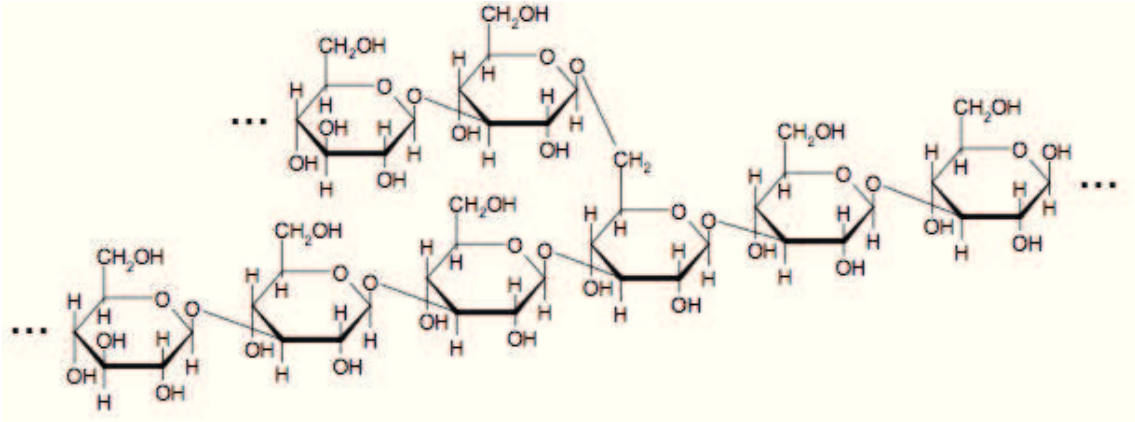


Figure 2. Schematic of a laminarin fragment [44].

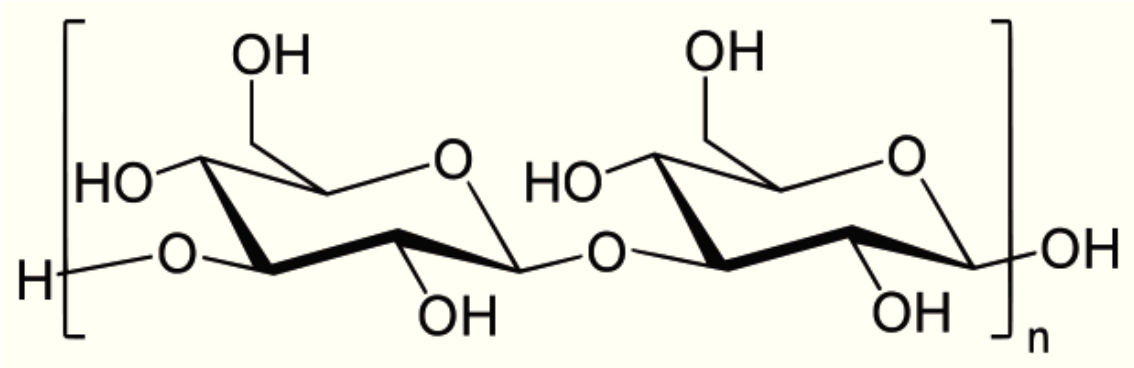


Figure 3. Schematic of a callose dimer [45].

In fact, chrysolaminarin is not the only □-1,3-glucan in diatoms. Callose is another important □-1,3-glucan but serves as a structural molecule. It is a linear molecule formed by □-1,3-linked glucose. Callose forms the so-called gasket or joint that seals the epitheca and hypotheca [42]. Because the frustules are made of silica, which is resistant to hydrolysis, the degradation of callose is likely to be the first step towards the complete degradation of a diatom cell.

1.3 Carbohydrate-active enzymes (CAZymes)

The formation of glycosidic bonds is catalyzed by glycosyltransferases (GT), whereas their breakup is catalyzed by glycoside hydrolases (GH), polysaccharide lyases (PL) or carbohydrate esterases (CE) [46]. Together they are subsumed under the term

carbohydrate-active enzymes (CAZymes). These CAZymes are found in all domains of life and are responsible for the carbohydrate turnover inside living organisms [3]. Most CAZymes possess one catalytic module, but some have more than one modules to facilitate more complex reactions [1, 47]. It is noteworthy that some of these enzymes need carbohydrate-binding modules (CBMs) to recognize and bind their sugar substrates [48]. Information on these enzymes and modules is constantly collected by the Carbohydrate-Active enZyme database CAZy [46]. Given the heterogeneous of their substrate, CAZymes are potentially highly diverse. Different CAZymes have highly divergent sequences, folds, kinetics and catalytic mechanisms [46, 49].

Unlike Enzyme Commission (EC) numbers, the CAZy classification is not based on functional but on sequence and structural similarities [46, 49]. This classification scheme is convenient for studying the phylogenies of enzymes and their families. Also, members from the same CAZyme family may share the same catalytic kinetics and mechanisms [50], even though their substrates may be different [2]. However, it is sometimes difficult to predict the substrates and even the reactions of unknown members based on CAZyme classifications alone [46, 51]. Still, the CAZy database is an invaluable resource for the study of carbohydrate metabolism and it will continue to play a major role in the fast development of the modern glycobiology.

About 1 to 2% of the ORFs in an average genome are CAZymes [1-4]. Notably, some *B*□□□□□□□□□□□□ such as *B*□□□□□□□□□□□□ □□□□□□□□□□□□ □□□□□ VPI-5482 [52, 53] and *B*□□□□□□□□□□□□ □□□□□□□□□□□□ XB1A [54] even have over 7% of their genes annotated as CAZymes. GTs transfer sugar moieties from donors to acceptors and hence are responsible for the biosynthesis of polysaccharides and glycoconjugates in cells. Currently, there are 91 valid GT families in the CAZy database. Among those, GT5 members are involved in the syntheses of □-glucans such as glycogen and starch in cells, while GT35 members are glycogen phosphorylases involved in the degradation of □-glucans (Table 1).

Table 1. Examples ofCAZyme families and functions

CAZyme	Selected members	Selected functions
GT5	Glycogen glucosyltransferase	Syntheses of □-glucans
GT35	Glycogen phosphorylase	Lysis of glycogen
GH13	□ -Amylase	Degradation of □-glucans
GH13	Amylosucrase	Synthesis of external □-glucans
GH16	Laminarinase, licheninase	Degradation of □-1,3(4)-glucans
GH31	□-Glucosidase	Degradation of □-glucans

Glycoside hydrolases are found across all domains of life except some *A*□□□□□□ and a few unicellular parasitic eukaryotes [46]. The CAZy database currently features 131 GH families. At least three of these GH families contain □-glucan degrading enzymes, namely GH13, GH31 and GH57. GH13 contains □-amylases that breaks the □-glucan chain at random locations. In contrast, the □-glucosidases from GH31 are exo-glucanases that release a single glucose molecule per reaction (Table 1). GH57 members are thermostable □-glucanases that are widely found in the genomes of thermophilic bacteria and archaea.

The family GH16 also deserves special attention (Table 1). This family is one of the only eight GH families that are targeted at marine polysaccharides (GH16, 50, 82, 86, 96, 105, 107, 117 and 118). In summary, GH16 includes xyloglucan transglucosylases/hydrolases (XTHs), 1,3-□-galactanases, 1,4-□-galactanases/□-carrageenases, □nonspecific□1,3/1,3;1,4-□-D-glucan endohydrolases, and 1,3;1,4-□-D-glucan endohydrolases [55]. The substrates of these enzymes include agar (red algae), porphyran (red algae), laminarin (brown algae), chrysolaminarin (diatoms), □-carrageenan (red algae). There is currently no indication that GH16 can degrade pure □-1,4-cellulose. GH16 enzymes are found abundantly in marine bacteria, especially in the well-characterized *F*□□□□□□□□□□ like □□□□□□ □□□□□□□□□□□□□□ Dsij^{T} [56]. GH16 genes are also often found in clusters with susD and TonB-dependent receptor genes [31]. These clusters are often recognized as polysaccharide-utilization loci (PUL) in genomes and they are considered to be operons or regulons [57].

1.4 Enzymatic synergisms of CAZymes

Synergism of enzymes is defined as the cooperative enhancement of different types of enzymes acting together [58]. Synergistic behavior is best understood among cellulases. There are at least three types of enzymes engaging in the hydrolysis of cellulose microfibrils. Endoglucanses cut cellulose at random locations, exoglucanases are chain-end-specific and release cellobioses, and □-glucosidases split cellobiose disaccharides into glucose monomers. The degradation of cellulose is most efficient when these three types of enzymes cooperate [59, 60]. The mechanisms by which the synergism works are not quite clear. Currently, there is no evidence suggesting that enzyme-enzyme interaction is necessary [58]. So far, three other hypothetical mechanisms have been proposed. The first hypothesis states that the endoglucanases accelerate the downstream processes by cutting open the microfibril and generating new chain ends for exoglucanases. However, this is considered to be an oversimplification of a complex system [58, 61]. The second hypothesis is based on the observation that exoglucanases cannot pass through the amorphous regions of

cellulose. It suggests that it is the endoglucanases that degrade the amorphous regions so that exoglucanases can proceed [61]. But these two hypotheses may not be mutually exclusive. The first mechanism may become significant once the enzyme to substrate ratios are high [61]. The third one suggests that the exoglucanases can relieve product inhibition for the endoglucanases [62, 63].

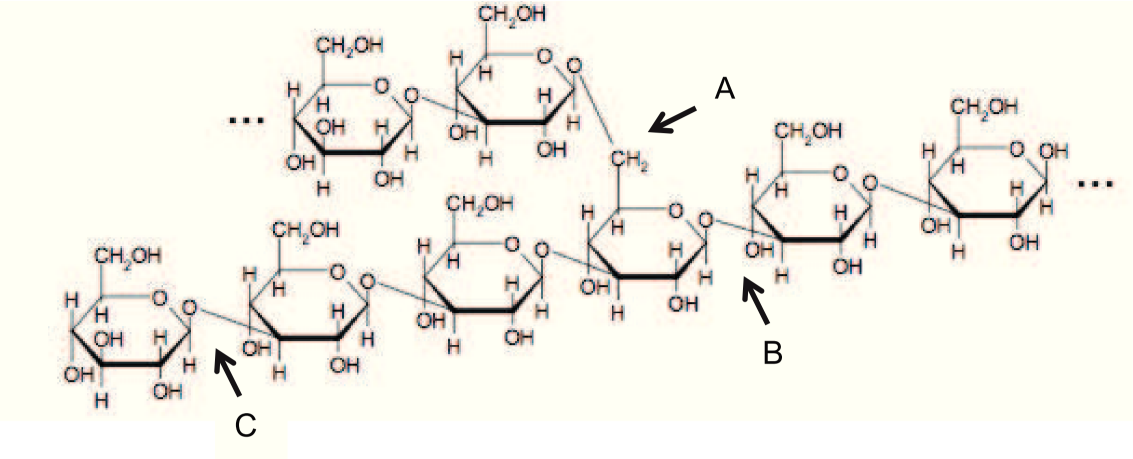


Figure 4. Schematic of the synergistic degradation of a chrysolaminarin molecule by A: β-1,6-glucanases; B: endo-□-1,3-glucanases and C: □-1,3-glucosidases.

A synergistic behavior of CAZymes was also observed in the degradation of laminarin by Pyrococcus furiosus [63]. Laminarin is a branched polymer. Similar to cellulose degradation, glucosidases can degrade the linear portions but they stall in the vicinity of the branching points. In laminarin, the linear portion is joined by β-1,3 bonds and the branches start with β-1,6 bonds. Few enzymes have both β-1,3- and β-1,6-glucanase activities. The rare exceptions include Gluc131A from Podospora anserina S mat+ [64] and FvBGL1 from the fruiting body of enoki mushrooms (*F*lammulina velutipes) [65]. In most cases, laminarin degradation is a joint operation of GH30 β-1,6-glucanases (EC 3.2.1.75) and □-1,3-glucanases. The latter has two distinct types, namely endo-acting forms from GH16 and GH64 (EC 3.2.1.39) and the exo-acting form from GH3, GH5 and GH17 (EC 3.2.1.58).

The synergism of CAZymes is a constant reminder of the complexity of polysaccharide degradation. As far as reaction efficiency is concerned, the combination of all types of enzymes is the most effective way [59, 63], in other words, the whole is greater than the sum of its parts. Also, to characterize a polysaccharide degrader and its substrates, it is recommended to account for all its hydrolase types. So far, the reaction type of the hydrolase can be predicted in silico by annotation and 3D structure prediction. In this thesis, it is hypothesized that *F*ormosa and Polaribacter are able to degrade laminarin. Hence, their genomic and metagenomic sequences were under thorough examination of endo- and exoglucanases with both approaches.

1.5 Trident: a bioinformatic pipeline for automatic CAZyme annotation

Since the establishment of the CAZy database, glycobiologists have never stopped searching for effective and accurate ways to identify and annotate CAZyme sequences. Their methods range from manual expert annotation to automatic computer annotation with varying success. Manual expert annotation starts by comparing the translated sequences against protein databases such as Swiss-Prot and ends up in cross-referencing the hit results with the CAZy database. For example, CAZymes in the genome of □□□□□□□ □□□□□□□□□□□□□□ Dsij^{T} [31] were annotated manually. Such an approach ensures a high accuracy of the annotations, but it trades in speed that is particularly important for large-scale next-generation sequencing data.

Automatic methods take advantage of the vast calculation capacities of modern computers and deliver results in only a fraction of the time needed for manual annotation. Although Needleman-Wunsch [66] and Smith-Waterman [67] algorithms remain the standard ways of finding perfect pairwise alignments, they just consume too much time for large amounts of long sequences. Since the 1990s, bioinformaticians have devised several algorithms for finding homologues against large databases in reasonable time. BLAST [68] is a landmark achievement in computational sequence alignment. It uses heuristics to cut the resources needed for fast alignments. Another widely used alignment method is based on profile Hidden Markov Models (profile HMMs) [69, 70]. In this approach, similar sequences are organized into a protein family. The alignment of each protein family is mathematically described by a profile. The profile is constructed based on Hidden Markov Model theory and represents the characteristics of the alignment. Profiles are then used in the search for new family members or to aid further sequence alignments. Compared to BLAST, the profile HMM approach uses position-specific scoring. In other words, it captures more sequence information than BLAST does [71]. For this reason, profile HMMs are considered to be more sensitive and more likely to find distant homologues than BLAST [72]. Profile HMMs are implemented in the software package HMMER [70] and the protein families are collected in the Pfam database [73-76]. However, HMMER does have one disadvantage against BLAST. Query sequences can only be identified if they belong to known protein families or domains and this is not always the case. In the face of large amounts of unknown genes from massive environmental sequencing, profile HMMs sometimes have poor identification rates.

Both BLAST and HMMER have been used to align CAZymes in genomes and metagenomes [77-82]. The results were then filtered according to a set of predefined rules, such as E-value or bit score cutoffs. However, even with stringent cutoff settings,

automatic annotation still ignores some important classification criteria. For example, the catalytic site is not always present in a short sequence but the latter can still have good E-value and bit score as long as the rest of the sequence aligns well with the reference sequence. In this case, automatic methods will consider it as a positive hit whereas a manual expert annotator will reject the result. Furthermore, multi-domain CAZyme sequences pose a challenge to computer algorithms, because these sequences can confuse algorithms in finding consensus annotations.

Because the next-generation sequencing technologies such as Illumina can generate hundreds of gigabases per run, such data can contain thousands of CAZyme genes, given that most metagenomic studies aim at fast understandings of the chosen communities and not the exact amounts of every single genes, it is not practical or necessary to manually annotate every single CAZyme sequence. Automatic annotation is in this case the only method of choice. Bearing in mind its drawbacks, automatic annotation should be subjected to stringent criteria to avoid false positives. After automatic annotation is finished, interesting candidate genes can be singled out and studied manually. For example, candidate CAZyme sequences can be validated by examining their catalytic sites provided by Prosite [83]. However, only 30 CAZyme families have their signature patterns listed in Prosite. Also, it is possible to get preliminary 3D structures, e.g. using high-performance prediction tools such as Phyre2 [84]. The Phyre2 prediction can provide useful information such as the folds, the catalytic mechanisms and the substrate-binding sites of the enzymes.

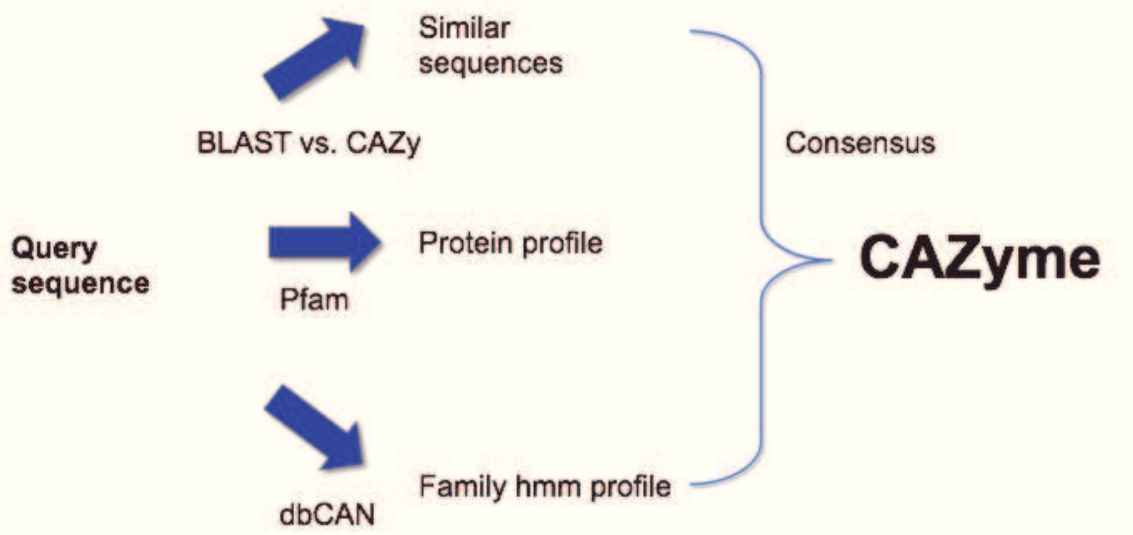


Figure 5. Workflow of the bioinformatic CAZyme annotation pipeline Trident.

A bioinformatic CAZyme annotation pipeline has been developed in this thesis. Considering the advantages and disadvantages of both manual and automatic

annotations, and knowing that no single algorithm can perfectly accomplish the task, this pipeline was designed to strike a balance between sensitivity and specificity. The pipeline Trident combines three automatic annotation approaches and computes consensus annotations for each sequence. Trident was used in Teeling [15]. It proves to be a fast and sensitive way to narrow thousands of ORFs down to several hundred CAZyme genes.

Trident consists of three annotation tools and one consensus-building tool. The first annotation tool is BLAST against the CAZyme sequences. CAZyme sequences were collected from the CAZy website [46] on a regular basis and formatted for BLAST. Hit results were only taken if the flexible E-values are lower than E-30. The second annotation tool is HMMER against the Pfam database and the results are semantically linked to CAZymes via rules defined either by CAZy or by Park [80]. The third tool is HMMER against the dbCAN database [81]. The two HMMER-based methods are subjected to a flexible E-value cutoff of E-15. The consensus-building algorithm takes all resulting hits into consideration. It applies either the simple "best hit" or the "majority rule" logic to give the final result.

Trident was calibrated and adjusted by benchmarking. Trident searched CAZymes in several public genomes such as Dsij^T [31]. The results were compared with those posted on the CAZy website. Trident generally predicts more CAZymes than those on the website. Usually they resulted from multi-domain CAZymes like GH5. Also, some sequences were wrongly classified as GH1/GH95 judged by that they do not contain the catalytic sites, although they well aligned to the putative sequences from and passed all the filters. This means that the results still need to be manually curated for accuracy.

1.6 CAZyme profiling can reveal an organism's lifestyle

The annotation of a single CAZyme can reveal useful information, such as its phylogenetic affiliation, catalytic mechanism and even its substrate specificity. Based on this kind of information, together with operon structure, signal peptide and transmembrane annotation, enzymatic models of polysaccharide utilization can be built and studied. In such models, several enzymes related to the utilization of a complex substrate are conceptualized as a molecular pipeline. The genes of these enzymes are often arranged in PULs. In such models, the subcellular localizations of the enzymes are considered, namely extracellular, transmembrane or intracellular compartments, as well as the their functions in the substrate flow. Despite of being conceptual, these models are invaluable for our understanding of the organism because they put relevant

enzymes into a cellular context and present the molecular machinery in a systematic way. Comparing these molecular pipelines from different genomes can often reveal the evolutionary relations, ecological niches and life styles of organisms.

The detailed characterization of PULs is one application of CAZyme analyses in genome studies. It can unravel how the organism enzymatically processes a specific substrate in great detail. However, this approach can only depict one aspect of the organism at a time. Another approach, termed CAZyme profiling, can reveal genome-wide CAZyme capacities at a glance.

A CAZyme profile is a tally of all the CAZyme genes in a target genome. In its simplest form, only the gene number of each CAZyme family is shown without any annotation details. Such a summary of CAZyme families alone can sometimes reveal useful information that otherwise can only be obtained through more sophisticated methods. For example, a comparison between the CAZyme profiles of *C*□□□□□□□□ Pelagibacter ubique HTCC1062 and □□□□□□□ □□□□□□□□□□□□□□□ Dsij^{T} can already demonstrate the difference between their lifestyles. P. ubique HTCC1062 has a small repertoire of GH and GT families and each of them has just a small numbers of genes. Among the four GH families, GH23, GH73 and GH103 are involved in bacterial cell wall turnover. It is obvious that P. ubique HTCC1062 does not have much polysaccharide degradation capacity. In fact, this collection of enzymes could represent the minimal CAZymes for a free-living organism. Furthermore, this CAZyme profile also reflects □. □□□□□□ HTCC1062□s small-size genome and oligotrophic lifestyle. This minimal dependence on external polysaccharides is one of the reasons why P. ubique HTCC1062 is ubiquitously distributed in the global oceans. In comparison, the □. □□□□□□□□□□□□□□□ genome harbors 40 GH families. Among them, families GH16 and GH117 are involved in the degradation of agar [56], an important polysaccharide found in marine algae. Together with other □-glucanases that degrade alginate and □-carrageenan, □. □□□□□□□□□□□□□□□ is obviously an algae-degrader. On the other hand, this algae-dependency limits the distribution of □. □□□□□□□□□□□□□□□. Indeed, □. □□□□□□□□□□□□□□□ is less frequently found as indicated by the marine metagenomes discussed later in this thesis.

Besides pairwise comparison of CAZyme profiles, it is also possible to look at the relations among multiple CAZyme profiles all at once. The relations of different CAZyme profiles can be measured by their Euclidean distances. The Euclidean distance is defined as the root-sum-square of difference in each family between two CAZyme profiles:

$$d(q,p) = \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

(In this formula, q_i and p_i are gene counts of profile q and p in CAZyme family i.)

With this formula, the distances of one CAZyme profile towards all the others in the group can be calculated. These distance values can be sorted and the so-called "nearest neighbors" can be identified. These nearest neighbors can sometimes provide useful insights into the target genome. For example, the genome of *G* KT0803 [85] shows that it shares a rather similar CAZyme profile with the two *F* genomes of *C* DSM 14237 [86] and *C* DSM 7489 [87]. The presences of GH117 in all three profiles suggest that they are all capable of degrading agar. Indeed, *C.* and *C.* are known to be agarolytic [87, 88], while this has yet to be confirmed for *G.* Also among other similar features, all these genomes possess GH43 that indicates their abilities of xylan degradation. Finally, both *C.* and *C.* are capable of producing a wide range of extracellular enzymes to degrade polysaccharides [86-88]. The same feature was implied in the description of *G.* s marine free-living lifestyle [85]. Therefore, the close distances among these three bacteria coincide with their similar lifestyles. When little is known about a genome, its nearest neighbors can provide hints about its living strategy.

However, caution is needed to interpret the results. Because there is no hard cutoff or criterion, it is sometimes hard to define which distance can be considered biologically meaningful. And since every pair of CAZyme profiles has a Euclidean distance ranging from zero to infinitively large, its value can only be interpreted through comparisons. Also, when we interpret the results, critical considerations are needed to distinguish biological meaningful hypotheses from those that are not.

1.7 Combined taxonomic classification and CAZyme profiling can reveal trophic relations in metagenomes

1.7.1 Taxonomic classification of metagenomes

The majority of environmental bacteria are non-cultivable with current methods [89] and this greatly limits our understanding of microbial ecology. An approach to circumvent this limitation is to sequence the environmental samples directly without isolation and cultivation. This is called metagenomics [90, 91]. With the advent of high-

throughput and low-cost sequencing technologies, metagenomic studies are generating ever-larger amounts of sequencing data. These data are revolutionizing our view on the environment. They have revealed large amounts of previous unknown organisms and novel genes. For ecologists, metagenomics is a fast way to gain an overview of some highly complex microbial communities. It is no wonder that metagenomics is getting more and more popular across microbiology research communities.

CAZyme profiling can be applied to metagenomes, too. The first obvious approach is to profile the whole metagenome as if it was one big single genome. The resulting CAZyme profile hence represents the carbohydrate metabolism of the total community. This approach can quickly reveal the potential of microbial communities and possible interactions with their habitats. Two exemplary studies, the □□□ □ □□ □ □□□□□ foregut microbiome by Pope □□ □□ [92] and the leaf-cutter ant fungus garden microbiome by Suen □□□□ [93] have showcased the power of this approach.

However, metagenomes have much more to offer. Metagenomes contain individual DNA fragments from organisms that interact with each other in the same place at the same time. Genes that make all these interactions possible are captured in the metagenomes. Detailed annotations of these genes can indicate their biochemical reactions, reactants and products. These metabolic products can be transferred among different organisms in food webs. For each metabolite transferred between organisms, the producer has the synthetic genes and the consumer has the degradation ones. If there is a way to first separate the genes between producer and consumer, it is then possible to identify their complemented gene sets and even their shared metabolite. That involves two techniques in metagenomics □ taxonomic classification of metagenome sequences and functional annotation of their metabolic genes. And the latter includes CAZyme profiling.

Taxonomic classification of metagenome sequences, especially short sequences, is one of the most challenging tasks in metagenomics. These sequences contain incomplete taxonomic signals and they also often come from as yet unknown organisms. To assign these sequences to established taxonomic groups, both the intrinsic and the extrinsic classifications were developed.

Intrinsic tools rely on the information within the sequences alone and no external data is involved. This intrinsic information is also understood as DNA signatures, such as tetranucleotide frequencies used by TETRA [94] and TaxSOM [95] (first manuscript; Chapter 2.1). Other notable intrinsic tools include the na□ve Bayesian classifier [96], Phylopythia [97], Phymm [98] and TACOA [99]. In essence, they compress the sequence information into a set of values and bin sequences with similar values.

Sequences from the same bin are considered to be taxonomically related. The taxonomic affiliations of any unknown sequences can therefore be inferred from the other known group members. Prior to the classification, the tools have to be trained to recognize the DNA signatures. Although the training processes can be lengthy and computation-intensive, the classification per se is fast. However, DNA signatures such as tetranucleotide frequencies discard the position-dependent information of the nucleotides and thus are considered to be weak signals. Generally, sequences with sufficient lengths are required to generate reliable signatures [100].

In contrast, extrinsic tools compare the query sequences against databases. If hits are found after the database search, a reasoning process consolidates the results and assigns the query sequence to a taxon. Compared to their intrinsic counterparts, extrinsic tools do not require extensive training before use. Also, they are less demanding in terms of sequence length [100]. In the preparatory phase of this study, three extrinsic tools were developed: a modified DarkHorse algorithm [15, 101, 102], an algorithm named Kirsten (kinship relationship re-establishment) [15, 102] and CARMA [15, 72, 102]. The first two tools are based on BLAST results while the third is Pfam-based. They evaluate the taxonomic information in the hits but use different logics to draw their conclusions. In detail, DarkHorse and Kirsten collect the top ten BLAST results if their E-values are lower than a predefined threshold. Afterwards, the taxonomic information of each hit gets expanded to the 29 taxonomic ranks defined by NCBI. DarkHorse calculates the occurrence of each taxonomic term and chooses the one with the highest frequency score. Kirsten, in contrast, takes the bit scores into consideration and carries out rank-specific evaluations. The frequency score is calculated as the bit score sum of each taxonomic term. If the highest frequency score on a particular rank is lower than a predefined threshold, the evaluation stops. CARMA compares query sequences against Pfam. If the searches are successful, the query sequences are then pooled with close reference sequences. Afterwards, neighbor-joining trees are constructed in order to identify the closest neighbors for the query sequences. The taxonomic affiliations of the queries can finally be determined after Kirsten runs on their close neighbors.

As already mentioned in the previous section, taxonomic classification of short sequences is a daunting task and is prone to errors. To reduce the amount of misclassification, it is recommended to combine results from diverse tools of different mechanisms. The consolidation of results is the task of the so-called meta-tools, such as Taxometer [15, 100], which has also used in the studies in this thesis.

1.7.2 CAZyme analysis of taxonomically classified sequences

After the sequences are taxonomically classified, they can be grouped based on their taxonomic affiliations. This process has been referred to as "taxobinning" and the sequence groups are "taxobins" [100]. The sequences of a taxobin represent the total genomic content of a particular taxon inside a metagenome. Taxobins separate sequences from different taxa and are the basis of studying their interactions. The amounts and frequencies of various functional genes can be calculated within each taxobin. In this way, we can generate CAZyme profiles for these taxobins and interpret them analogous to single genomes. Complemented with results of other annotation tools such as SignalP [103], TMHMM [104], Pfam [75] and BLAST [68], it is possible to study the sugar metabolisms of abundant taxa in a metagenome in depth, in particular when combined with expression analyses such as metatranscriptomics and metaproteomics. Given sufficient details, it is even possible to reconstruct the partial food web of the sample. Two metagenomic projects were analyzed this way in this thesis – the MIMAS project and the Logatchev metagenome.

1.8 Metagenomes from the Microbial Interactions in Marine Systems (MIMAS) Project

The Microbial Interactions in Marine Systems Project is an ongoing study of microbes in the North Sea, with a focus on bacterial carbohydrate degradation during and after spring phytoplankton blooms [15] (second manuscript; Chapter 2.2). The Helgoland Roads Project of the Biological Institute Helgoland (BAH) provides the long-term physiochemical data such as turbidity, temperature, salinity and the concentrations of silicate, phosphate, nitrate, nitrite and ammonia. The [illegible]-[illegible] cell abundances were measured via Catalyzed reporter deposition Fluorescence [illegible]-[illegible] Hybridization (CARD-FISH) [105] twice a week. Samples from several dates were subjected to direct DNA and cDNA 454 pyrosequencing, 16S rRNA pyrotag sequencing and metaproteome analyses. These data captured the onset, development and termination of a diatom spring bloom and its subsequent bacterioplankton bloom in 2009, as well as the changes of the environmental parameters. As indicated by the chlorophyll a data, the algal bloom started around the 2nd of March, reached its maximum around March 23rd and gradually decreased towards the end of April. The diatom population was dominated by centric [illegible] spp. [15]. The diatom bloom apparently induced a succession of blooms of distinct bacterioplankton clades. The first bacterial bloom occurred from 20/03/09 to 09/04/2009 and it consisted mainly of *F*[illegible] sp. (class

F□□□□□□□□□□□□). Their relative abundances rose quickly from below 1% to 23% of the bacterial population with cell counts increasing from 7.8 □ 10^2 to 3.4 □ 10^5 cells/ml. Afterwards, the numbers of *F*□□□ □□□ decreased sharply below the CARD-FISH detection limit. The genera □□□□□□□ (class: *G*□□ □ □□□□□□□□□□□□□□□□) and □□□□□□□□□ (class *F*□□□□□□□□□□□□) started to bloom one week later. □□□□□□□ relative abundances increased from 0.2% to 16% (1.6 □ 10^3 to 1.5 □ 10^5 cells/ml) but dropped below 1% after two weeks. At the same time, □□□□□□□□□□relative abundances rose from 10% to 20% (8.0 □ 10^4 to 1.5 □ 10^5 cells/ml). Several CAZyme families, sulfatases, and sugar transporters were identified in the blooming bacterioplankton using a combination of metagenome and metaproteome analyses, but the substrates or metabolites remained elusive [15].

Although this microbial succession was described in detail, the machinery behind such a succession has yet to be clarified. Two possible explanations were considered. The first one is the so-called □top-down□control. Under this hypothesis, the abundance of microbes was controlled mainly by predators [106]. The rapid increase of microbes could induce the growth of predators such as heterotrophic nanoflagellates [106, 107] or alternatively the outburst of viruses [108]. This higher mortality would terminate the blooms of particular clades. The second hypothesis is the so-called □bottom-up□control and assumes that the limiting factor in the bloom was nutrient availability, not mortality [15]. Although these two hypotheses examine the succession from two different angles, in reality however, the observed phenomenon was more likely to be the result of both top-down and bottom-up mechanisms.

The MIMAS study was focused on the metabolic capacities of microbes and indeed found evidences of bottom-up control and nutrient niche partitioning. Metagenome taxobins from different time points not only showed distinct CAZyme profiles, but also different transporters profiles and sulfatase numbers. The shift in the gene repertoires reflected changes of substrates availability. The whole chain of events could in essence be considered driven by a succession of substrates. In order to shed more lights onto this hypothesis, a more detailed study is necessary to address several key questions. First, what were the polysaccharide degradation capacities of the key players? Were they able to degrade the same kind of substrates or were they rather specialists? Second, what were the substrates? Third, were there any trophic connections among them? Finally, why did the three *F*□□□□□□□□□□□□□□□ appear at different times during the algal bloom?

The third manuscript of this thesis (chapter 2.3) attempts to answer these questions based on a more in-depth analysis of the metagenome data. This analysis added further detail to the CAZyme analyses published in the previous study [15]. The

abundant taxa at different sampling time points possessed contrasting CAZymes, transporter and sulfatase profiles. These different profiles also suggest that these key players did not have a uniform degradation spectrum. A further investigation of the substrate specificities of CAZymes has even indicated polysaccharides that might have been transmitted through trophic levels. The diatoms produced large amounts of sulfated extracellular polysaccharides [109] and intracellular □-1,3-glucans [110]. *F*ormosa possessed the exact degradation enzymes for them – large numbers of sulfatases and □-1,3-glucanases from GH16 and GH17. For this reason, the genus *F*ormosa was considered to be the primary diatom degrader. Furthermore, the *F*ormosa taxobin appeared to be able to synthesize □-glucans, both extracellularly by amylosucrases from GH13, and intracellularly by glycogen synthase from GT5. The second wave of bacterioplankton consisted of Reinekea and Polaribacter These two bacteria contained □-glucanases from GH13 and □-glucosidases from GH31, which indicated that they were able to degrade the □-glucan probably of *F*ormosa origin. In addition, Polaribacter also had CAZymes that degrade chrysolaminarin and agar, two prominent algal polysaccharides. Based on these findings, it is possible to reconstruct a partial food web of the event. The *F*ormosa was involved in the primary degradation of the diatom cells, while Reinekea and Polaribacter were able to feed on in parts the □-glucans from *F*ormosa . Additionally, Polaribacter could also degrade the chrysolaminarin present within the diatom cells as a storage compound. This study represents an application of CAZyme profiling in a real-world metagenome project and demonstrates the strength of this method: the generation of fine-granular hypotheses about the life styles of particular clades. Fortunately, the resolution was good even on the genus level in the MIMAS Project. However, such a success is only possible given enough sequencing coverage and supporting data. In the MIMAS study, samples were taken from the surface seawater relatively easily. For the metagenomic analyses, several samples are sequenced with at least two PTPs. For the genomes of the key players such as Thalassiosira , *F*ormosa , Reinekea and Polaribacter , not only public data are available, but also six new draft genomes were sequenced within the project. Furthermore, the physiologies of T. pseudonana , *F.* agariphila , R. blandensis and P. irgensii were well studied, all of which provide a vital starting point for the project.

1.9 Metagenomes from the Logatchev deep-sea hydrothermal vent field

Another CAZyme profiling of this doctoral thesis was done for a much more remote environment. The Logatchev hydrothermal field (LHF) is located at 15 □N the Mid-Atlantic Ridge [111]. It is hosted by ultramafic components composed of Earth mantle-

derived peridotite. LHF is characterized by vent fluids enriched in dissolved hydrogen and methane and thick sediments, sometimes covered with white mats [112]. The location has a water depth of 3,000 meters and is far beyond the reach of sunlight, yet there exist highly diverse microbial communities driven by chemical energy. Chemolithoautotrophic bacteria use hydrogen sulfide, methane and hydrogen as energy source. These bacteria form the basis of the food web in these ecosystems. In LHF, the majority of sulfur-oxidizing bacteria belong to the bacterial classes *G* and *E*. Together with the *D*, they composed the bulk of the LHF microbial community. In a metagenome where only 70% of reads could be classified as *B*, these bacterial classes accounted for 37% of all metagenomic reads. This also means that these three classes constitute the major part of the LHF microbial landscape. As a result, they hold the key to our understanding of this deep-sea ecosystem. One part of this thesis contributed a CAZyme profiling to a manuscript dedicated to the metagenomic characterization of these three classes (fourth manuscript; Chapter 2.4). The metabolic potential of the proteobacterial class for carbon, sulfur and nitrogen cycling was investigated. In this context it is noteworthy that studies of hydrothermal vent habitats so far have focused on bacterial primary production and for the most part neglected follow-up heterotrophic substrate conversions. CAZyme profiles can on the one hand provide hints about unique metabolic potentials of the studied deep-sea bacteria and on the other hand may reveal pathways and substrates that are common between surface water and deep-sea habitats.

1.10 The aims of this thesis: establishment and application of CAZyme profiling for studies in microbial ecology

The major aim of this doctoral thesis was to add CAZyme profiling as a powerful tool to the metagenomic toolkit. Nowadays, microbial genomic or metagenomic studies have already gone far beyond the simple characterization of genes. They are providing insights into the interactions among the organisms and even the evolution history [82]. Microbiologists can learn and draw conclusions about the abiotic and biotic environment through the genomes. CAZymes, on the one hand, reflect microbe's potential of metabolizing certain kinds of carbohydrates. On the other hand, they may also pinpoint microbe's position in a food web.

Furthermore, extremophilic CAZymes are of high industrial interest because they remain functional under a wide range of physiological conditions including temperature, NaCl concentration and pH [113]. For example, salt-tolerant xylanases have potentials

in food possessing under high-salt conditions [114, 115], but they are rarely researched [116]. GH43 xylanases from *H*□□□□□□□□□□ □□□□ □□□□ SARL4B^T are adapted to a high-ionic environment because of the □salt-in□ strategy of their host (see also fifth manuscript; chapter 2.5). These xylanases are currently under a screening process and awaiting further characterizations.

This study is not the first one in which CAZyme profiling was used in an ecological study. The leaf-cutter ant fungus gardens study showed how the fungus garden microbes specialize in degrading high plant biomass [93]. CAZyme profiling confirmed that the human gut microbiome contains plant cell wall degradation enzymes that lend human the ability to digest plant fibers [78]. It was rather my aim to further establish CAZyme profiling as a tool for the investigation of marine food webs. In this study, CAZyme profiling was applied to several habitats. Its application contributed significantly to these studies of marine microbial ecology.

2. P□b□ica□i□□□ a□d Ma□□□c□i□□□

During my doctoral thesis I contributed to eight manuscripts (Table 2). The abstracts of three of these manuscripts have been moved to the Appendix, because first of all, they are still in an early draft status and secondly, for better clarifying the focus of my doctoral thesis. In this section, I listed five manuscripts in which I was responsible for the development and application of a bioinformatic pipeline for gene prediction, taxonomic classification and functional annotation of CAZymes. Based on CAZyme profiling, a food web was reconstructed in the MIMAS study, a possible evolutionary scenario was proposed for *H*□□□□□□□□□□ □□□□ □□□□ SARL4B^T and two polysaccharide metabolisms were confirmed in the deep-sea habitat Logatchev hydrothermal vent.

Table 2. Overview of the manuscripts, to which this thesis contributed to.

Manuscript	Authors	Titles	Journal/Status
2.1	Weber □□ □□. 2011	Practical application of Self-Organizing Maps to interrelate biodiversity and functional data in NGS-based metagenomics	□□□ *I*□□ *E J*□□□□□□□
2.2	Teeling □□ □□. 2012	Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom	□□□□□□□
2.3	Huang □□□□.	Carbohydrate-active enzyme profiling of a diatom-induced bacterioplankton succession reveals niche adaptations and trophic connections	In preparation for BMC Genomics
2.4	Suenaga □□□□.	Metagenomics reveals niche-differentiation and habitat-specific adaptation in surface sediments of the Logatchev hydrothermal field	In preparation for Environmental Microbiology
2.5	Werner □□□□.	*H*□□□□□□□□□□ □□□□ □□□□: Complete genome sequencing and proteomics identify the first cultivated euryarchaeon from a deep-sea anoxic brine lake as polysaccharide degrader	In preparation for Environmental Microbiology
7.1	Jan □□□□.	The gill chamber epibiosis of deep-sea □ □□ □□□□□□ □□□□□□□□□□ shrimp thoroughly investigated by metagenomics and discovery of zetaproteobacterial epibionts	Submitted to □□□ *I*□□ *E J*□□□□□□□
7.2	Mann □□□□.	Complete genome sequence of the algae-associated marine flavobacterium *F*□□□ □□□ □□□□□□□□□ KMM 3901^T	In preparation for Journal of Bacteriology
7.3	Huang □□□□.	Geomicrobiology of Hot Lake, a shallow-sea hydrothermal vent site off Panarea Island, Italy	In preparation

2.1 Practical application of Self-Organizing Maps to interrelate biodiversity and functional data in NGS-based metagenomics

Authors

Marc Weber[1], Hanno Teeling[1]*, Sixing Huang[1], Jost Waldmann[1,2], Mariette Kassabgy[1], Bernhard M Fuchs[1], Anna Klindworth[1], Christine Klockow[1,3], Antje Wichels[4], Gunnar Gerdts[4], Rudolf Amann[1] and Frank Oliver Gl□ckner[1, 3]

[1] Max Planck Institute for Marine Microbiology, Bremen, Germany
[2] Institute of Marine Biotechnology e.V., Greifswald, Germany
[3] Jacobs University Bremen gGmbH, Bremen, Germany
[4] Alfred Wegener Institute for Polar and Marine Research, Biologische Anstalt Helgoland, Helgoland, Germany

*Corresponding author: Hanno Teeling

Publication status:

Published in □□□ *I*□□ *E J*□□□□□□

My contribution

TaxSOM is a taxonomic classification tool based on Self-Organizing Map. This paper explained the concept and implementation of TaxSOM, as well as its application in five metagenome datasets from the Microbial Interactions in Marine Systems (MIMAS) Project [15]. In order to evaluate TaxSOM, I implemented two protein-level taxonomic classification tools Kirsten and CARMA [72]. To benchmark the performance and validate the results of TaxSOM, I analyzed the same datasets with Kirsten and CARMA independently. The classification results of the tools were broken down into the domain, phylum, class, order, family and genus levels. We calculated the sensitivity and specificity of tools for comparison. We compared the metagenome biodiversity with the results of CARD-FISH. During the development of TaxSOM, I made improvement suggestions for the software. I also participated in the interpretation and discussion of the results.

The ISME Journal (2011) 5, 918–928

www.nature.com/ismej

ORIGINAL ARTICLE

Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics

Marc Weber[1], Hanno Teeling[1], Sixing Huang[1], Jost Waldmann[1,2], Mariette Kassabgy[1], Bernhard M Fuchs[1], Anna Klindworth[1], Christine Klockow[1,3], Antje Wichels[4], Gunnar Gerdts[4], Rudolf Amann[1] and Frank Oliver Glöckner[1,3]

[1]*Max Planck Institute for Marine Microbiology, Bremen, Germany;* [2]*Institute of Marine Biotechnology e.V., Greifswald, Germany;* [3]*Jacobs University Bremen gGmbH, Bremen, Germany and* [4]*Alfred Wegener Institute for Polar and Marine Research, Biologische Anstalt Helgoland, Helgoland, Germany*

Next-generation sequencing (NGS) technologies have enabled the application of broad-scale sequencing in microbial biodiversity and metagenome studies. Biodiversity is usually targeted by classifying 16S ribosomal RNA genes, while metagenomic approaches target metabolic genes. However, both approaches remain isolated, as long as the taxonomic and functional information cannot be interrelated. Techniques like self-organizing maps (SOMs) have been applied to cluster metagenomes into taxon-specific bins in order to link biodiversity with functions, but have not been applied to broad-scale NGS-based metagenomics yet. Here, we provide a novel implementation, demonstrate its potential and practicability, and provide a web-based service for public usage. Evaluation with published data sets mimicking varyingly complex habitats resulted into classification specificities and sensitivities of close to 100% to above 90% from phylum to genus level for assemblies exceeding 8 kb for low and medium complexity data. When applied to five real-world metagenomes of medium complexity from direct pyrosequencing of marine subsurface waters, classifications of assemblies above 2.5 kb were in good agreement with fluorescence *in situ* hybridizations, indicating that biodiversity was mostly retained within the metagenomes, and confirming high classification specificities. This was validated by two protein-based classifications (PBCs) methods. SOMs were able to retrieve the relevant taxa down to the genus level, while surpassing PBCs in resolution. In order to make the approach accessible to a broad audience, we implemented a feature-rich web-based SOM application named TaxSOM, which is freely available at http://www.megx.net/toolbox/taxsom. TaxSOM can classify reads or assemblies exceeding 2.5 kb with high accuracy and thus assists in linking biodiversity and functions in metagenome studies, which is a precondition to study microbial ecology in a holistic fashion.
The ISME Journal (2011) **5,** 918–928; doi:10.1038/ismej.2010.180; published online 16 December 2010
Subject Category: microbial ecology and functional diversity of natural habitats
Keywords: binning; metagenomics; molecular ecology; self-organizing map (SOM); taxonomic classification; TaxSOM

Introduction

The launch of next-generation sequencing (NGS) was nothing less than a paradigm shift in environmental molecular microbiology. The dramatic drop in sequencing costs that followed has resulted in an unprecedented rate of growth in microbial genome sequences. This development has spurred the establishment of sequencing initiatives aiming to explore the realm of microbial genomes in more targeted ways than before, for example, by focusing on specific habitats or taxa. For example, the 'Marine Microbiology Initiative' of the Gordon and Betty Moore foundation has contributed almost 200 draft genomes from marine habitats, and the 'Genomic Encyclopedia for Bacteria and Archaea' project of the Joint Genome Institute and the German Collection of Microorganisms and Cell Cultures (DSMZ) has begun to systematically fill the remaining gaps in the prokaryotic branches of the tree of life by aiming to sequence at least one representative from all clades (Wu *et al.*, 2009). The introduction of NGS has also propelled metagenomic community-sequencing approaches, which led to dedicated initiatives as well. For marine habitats, the 'International Census of Marine Microbes' is focusing on extending microbial biodiversity knowledge by the

Correspondence: H Teeling, Department of Molecular Ecology/Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, Bremen 28359, Germany.
E-mail: hteeling@mpi-bremen.de
Received 21 June 2010; revised 27 October 2010; accepted 27 October 2010; published online 16 December 2010

large-scale sequencing of 16S ribosomal RNA V6 hyper-variable regions (Sogin *et al.*, 2006; Huse *et al.*, 2008), while integration of the wealth of metagenomic data from different sources is at focus of the 'Community Cyberinfrastructure for Advanced Marine Microbial Ecology Research and Analysis' (Camera) project (Seshadri *et al.*, 2007). Similar data integration projects have been established in the field of medical microbiology, such as the NIH Human Microbiome Project (Peterson *et al.*, 2009). Only a few years after the introduction of NGS, our picture of the microbial world is already becoming much clearer.

In spite of the advancements in DNA sequencing, currently available technologies still restrict low-cost full genome sequencing to cultivable strains. This requirement severely limits the application of NGS technologies to microbial biodiversity studies, because only a minor fraction (typically $<1\%$) of the microbial species in a given habitat can be cultivated with current techniques (Amann *et al.*, 1995; Huber *et al.*, 2007). It is anticipated that progress in single-cell isolation techniques (Ochman, 2007) and single molecule sequencing (Gupta, 2008; Clarke *et al.*, 2009; Eid *et al.*, 2009) will soon overcome this limitation. For now, however, metagenomics the sequencing of DNA from an environmental sample without previous species separation or cultivation, is the method of choice for obtaining longer fragments from the genomes of the vast majority of as-yet uncultured microorganisms.

In the classical metagenome approach, genomic libraries are constructed by cloning fragmented environmental DNA into vectors that are subsequently amplified in ultra-competent host cells (Schloss and Handelsman, 2003). Once a metagenome library is constructed, it can be screened for inserts carrying specific genes or metabolic activities. These strategies have been termed sequence- and function-driven approaches (Schloss and Handelsman, 2003) and are used to select dedicated inserts from the library for full-length sequencing. These approaches, however, have the inherent disadvantage of limiting obtainable sequence information to the few genes adjacent to the respective target genes.

With the advent of NGS it has become feasible to omit the cloning step and sequence environmental DNA directly. In particular, if the target organism is in high abundance or even dominates a habitat, the sheer power of NGS allows for obtaining longer genomic fragments by direct sequencing and assembly of extracted environmental DNA. In contrast, direct DNA sequencing of habitats with high overall biodiversity or low-abundance target species mostly yields sequences harboring partial or single genes and relatively few longer assemblies with multiple genes. When a specific microorganism or function is desired, therefore, the classical metagenome approach is still much more favorable. If, however, community function in low to medium biodiverse habitats as a whole is at focus, then direct sequencing is a viable approach. Although brute-force direct sequencing of such microbial communities does not yield individual genomes, it often yields longer assemblies of the most abundant species and a wealth of sequences that can be taxonomically clustered into bins (taxobins) and subsequently mined for functions. This approach requires, of course, methods that allow these sequences to be taxonomically classified with reasonable accuracy.

In general, taxonomic classification of metagenomic DNA fragments can be achieved either on the level of the encoded genes or on the level of the DNA sequence themselves.

An introduction to gene-level taxonomic classification is beyond the scope of this article. In brief, they are either based on the post-processing of BLASTP (Altschul *et al.*, 1990) searches as in Phylogena (Hanekamp *et al.*, 2007) or MEGAN (Huson *et al.*, 2007), or on the post-processing of Pfam searches (Sonnhammer *et al.*, 1997) as in CARMA (Krause *et al.*, 2008).

Taxonomic classification of DNA sequences on the level of base composition is still unintuitive to many biologists. However, not only the genes but also the DNA itself—including non-coding regions—is subjected to various evolutionary forces (Karlin *et al.*, 1998), like species-specific codon preference, constraints because of DNA superstructure and $G+C$ content maintenance, and biases that are introduced by the replication machinery. As a result, DNA carries a fingerprint-like species-specific signature in its base composition that is most pronounced in the patterns of statistical over- and underrepresentation of short oligonucleotides from tetra- to hexanucleotides (McHardy *et al.*, 2007). As the factors that give rise to these fingerprints are inheritable, they also carry a detectable albeit weak phylogenetic signal (Pride *et al.*, 2003). The first work on genomic DNA signatures dates back to well before the genomic area started with the sequencing of the first complete bacterial genome (Fleischmann *et al.*, 1995) and was pioneered among others by Samuel Karlin *et al.* (Burge *et al.*, 1992). At first, scientists have investigated this phenomenon with rather simplistic methods like dinucleotide or tetranucleotide relative abundances (Karlin and Ladunga, 1994; Karlin *et al.*, 1994, 1998; Karlin and Burge, 1995; Karlin, 1998). Later, however, a whole variety of different methods have been applied to oligonucleotide signatures, such as Markov models (Rocha *et al.*, 1998; Pride *et al.*, 2003; Reva and Tümmler, 2004; Teeling *et al.*, 2004), frequency chaos game representations (Deschavanne *et al.*, 1999) and Bayesian classifiers (Sandberg *et al.*, 2001). More recently, machine-learning algorithms have been applied to the task. These can be subdivided into supervised algorithms like support vector machines, and unsupervised algorithms like kernelized nearest-neighbor approaches and self-organizing maps (SOMs). Support vector machines

have been used in PhyloPhythia (McHardy *et al.*, 2007), a kernelized nearest-neighbor approach in TACOA (Diaz *et al.*, 2009) and SOMs have been used in a variety of different variants, like batch-learning SOMs (BLSOMs) (Abe *et al.*, 2003, 2005), growing SOMs (GSOMs) (Chan *et al.*, 2008a, b), hyperbolic SOMs (Martin *et al.*, 2008) and emergent SOMs (Dick *et al.*, 2009). One of the most recent DNA-based approaches to taxonomically classifying metagenomic DNA fragments is the usage of interpolated context models, as implemented in Phymm and PhymmBL (Brady and Salzberg, 2009).

Here, we explore the practical application of a novel implementation of GSOMs and BLSOMs for the taxonomic classification of metagenome data sets. We first demonstrate the performance of both SOM variants on the basis of previously published simulated metagenomes (Mavromatis *et al.*, 2007) as well as data from complete microbial genomes. Then, we demonstrate how SOMs can be applied to real-world metagenomes for an overall taxonomic profiling as well as to follow community composition shifts over time. Our SOM implementation is termed TaxSOM and has been made available as a free and feature-rich web-service at http://www.megx.net/toolbox/taxsom.

Materials and methods

Implementation
TaxSOM has been implemented in the C++ programming language using the ocount2 (http://www.promedici.de/ocount2), Lapack++ (http://lapackpp.sourceforge.net), MySQL++ (http://tangentsoft.net/mysql++/) and Boost (http://www.boost.org) C++ libraries. Ocount2 has been used for oligonucleotide counting and Markov model-based z-transformations. Lapack++ has been used for Eigenvector transformation and other matrix operations for principal components analysis, MySQL++ for handling MySQL queries and Boost for parsing program options and serialization of computed SOMs. Boost Python libraries were used to provide an easy way for wrapping TaxSOMs C++ functions and make them accessible in the Python programming language. TaxSOM's web-interface was implemented in PHP (Hypertext Preprocessor) in conjunction with some Python scripts for data processing and with scalable vector graphics for SOM visualizations.

SOM specificity and sensitivities
Specificity (true positives/(true positives + false positives)) and sensitivity (true positives/(true positives + false negatives)) were used as classification accuracy measures. A classification was considered as a true positive, when a query sequence was classified on a SOM node representing only sequences of the query's taxonomic affiliation. It was considered as false positive, when a query sequence was classified on a node representing only sequences from different taxonomic affiliation. Classification of sequences that ought to be classified but were matching ambiguous nodes representing multiple taxa were treated as false negatives. In addition, the F-measure value, which is the harmonic mean of specificity and sensitivity, was used (see Supplementary Tables 1, 2, and 3).

Simulated metagenome data sets for evaluation
In order to evaluate the accuracy of TaxSOM's GSOM and BLSOM implementations for taxonomic DNA sequence classification, we used three previously published simulated data sets (simLC, simMC and simHC) of varying complexities (Mavromatis *et al.*, 2007). SimLC simulates a low-complexity community dominated by a single, near-clonal population that is flanked by low abundance species. SimMC was designed to mimic a moderately complex community like in the acid mine drainage biofilm (Tyson *et al.*, 2004) or the *Olavius algarvensis* symbionts' metagenome (Woyke *et al.*, 2006), wherein multiple dominant populations are flanked by low abundant ones. SimHC simulates a highly complex community with no dominant populations, like that present in agricultural soils (Tringe *et al.*, 2005). On all of these data sets, the three different assembly programs Arachne (Jaffe *et al.*, 2003), Phrap and JAZZ, have been used, resulting in a total of nine published test data sets. We excluded the JAZZ assemblies from our analysis, because they yielded a much lower number of correct taxonomic classifications than Phrap and Arachne assemblies. This is hence an effect of the JAZZ assembler (or its parameter settings) that would distort the subsequent taxonomic classification.

Data sets from known organisms for evaluation
A test data set comprising 1401 chromosomes and plasmids was constructed from all completely sequenced bacterial and archaeal genomes within GenBank. One-fifth was randomly cut from each sequence and retained for later classification, while the remaining 80% were used as training sequences for BLSOM and GSOM construction. A total of 10 SOMs were constructed, by splitting the training sequences into 10 or 50 kb fragments and using either di- tri- and tetranucleotide raw counts or z-scores as input data. The sequences remaining for classification were used to construct eight data sets of 0.5, 1, 2.5, 5, 10, 25, 30 and 50 kb lengths, which were subsequently classified by the SOMs (Supplementary Table 1).

Real-world data set
This study is part of the Microbial Interactions in Marine Systems project (MIMAS; http://www.mimas-project.de), which provided the real-world

Table 1 Real world metagenome data sets

Sampling date	*454 runs*	*Assembly*				
		No. reads	*No. contigs*	*Mb*	*Contigs >2.5 kb*	*Mb*
11 February 2009	1 591 182	2 PTP	56 160	31.7	227	0.8
31 March 2009	1 101 493	2 PTP	113 454	70.2	2321	9.8
07 April 2009	2 109 239	2 PTP	61 651	56.0	3229	15.5
14 April 2009	2 017 268	2 PTP	66 417	61.0	2999	16.2
16 June 2009	1 120 072	1 PTP	42 461	31.9	1137	5.0

Abbreviation: PTP, picotiter plate.

metagenome data (Table 1). The data consisted of pyrosequenced bacterial DNA from the coast off the North Sea Island Helgoland in the German bight (54° 11′ 3′′ N; 7° 54′ E) that was sampled at five different dates in 2009 (11 February, 31 March, 7 April, 14 April and 16 June).

At each of these points in time, 500 l of subsurface water (1 m depth) were sampled with the small research vessel Diker, immediately taken to the lab, and pre-filtered with 10 μm polycarbonate filters (TCTP, Millipore, Billerica, MA, USA) and 3 μm polycarbonate filters (TSTP, Millipore). The bacterial fraction was subsequently retained on 0.22 μm polyethersulfone filters (GPWP, Millipore). All filters were 142 mm in diameter and six membrane filtration units were operated in parallel to keep filtration times as low as possible. From the filters, bulk environmental DNA was extracted by a modified standard protocol (Zhou *et al.*, 1996). The DNA was then pyrosequenced directly on the GS FLX Ti platform with one (16 June) or two picotiter plates per sample (454 Life Sciences, Branford, CT, USA) by LGC Genomics (LGC Genomics GmbH, Berlin, Germany), and subsequently assembled with Newbler version 2.0.00.22 (Roche, 454 Life Sciences, Branford, CT, USA). From the assemblies, all sequences at least 2.5 kb long were taken for classification.

The bacterial community composition of the samples was assessed by catalyzed reporter deposition-fluorescence *in situ* hybridization (CARD-FISH) as follows: samples were fixed with 1% formaldehyde and 10 ml was filtered onto polycarbonate membrane filters (type GTTP, pore size 0.2 μm, Sartorius, Göttingen, Germany). CARD-FISH was performed according to previously published protocols (Pernthaler *et al.*, 2002). All hybridizations were counterstained with 4′,6-diamidino-2-phenylindole (1 μg ml^{-1}) and manually inspected and quantified.

Data sets for SOM construction

The SOMs for the evaluation of the simulated data sets were constructed from all bacterial and archaeal DNA sequences exceeding 485 kb (roughly the size of *Nanoarchaeum equitans*) in the NCBI GenBank database as of October 2008 (release no. 167). These sequences were extracted using a self-written C++ library termed phyloprint (Waldmann, 2008) that allows any type of sequence selection based on the complete NCBI taxonomy (phyloprint currently includes 462 019 nodes). This resulted in 1521 sequences comprising 3.43 Gb of DNA. All sequences were split into 50 kb fragments and subsequently used for the construction of GSOMs and BLSOMs with two types of inputs: oligonucleotide frequency raw counts (di-, tri- or tetranucleotide counts normalized on values between 0 and 1), or raw counts z-transformed based on a maximal order Markov model (Teeling *et al.*, 2004).

The real-world metagenome data sets were classified on a GSOM with tetranucleotide z-scores as input. The GSOM was trained in a habitat-specific manner using 340 bacterial and archaeal DNA genomic sequences from aquatic habitats, such as open ocean water, hot springs, hydrothermal vents or marine sediments. The respective habitat information was obtained from the EnvO-lite classifications present in the Marine Ecological Genomics (MEGX, http://www.megx.net/) database (Kottmann *et al.*, 2010), and the corresponding sequences were extracted from NCBI GenBank using phyloprint.

Protein-level taxonomic classification of real-world data sets for cross-evaluation

Protein-level taxonomic classification of the assembled 454-sequenced bulk environmental DNA was achieved as follows. First, the sequences were subjected to an open reading frame prediction with MetaGene (Noguchi *et al.*, 2006). Afterward, open reading frames exceeding 150 bp were compared with BLASTP (Altschul *et al.*, 1990) against the non-redundant NCBI database (as of 28 October 2008) and with hmmpfam (Eddy, 1996, 1998) against the Pfam database (release 22) (Sonnhammer *et al.*, 1997, 1998). Hits with good E-values (BLASTP: E ⩽E–15, hmmpfam: E ⩽E–5) were subsequently analyzed.

BLASTP hits were processed with an adaptation of the DarkHorse algorithm (Podell and Gaasterland, 2007). In brief, DarkHorse performs rank-based reasoning on the taxonomic terms from BLASTP hits, calculates for each hit a so-called lineage probability index and assigns the open reading

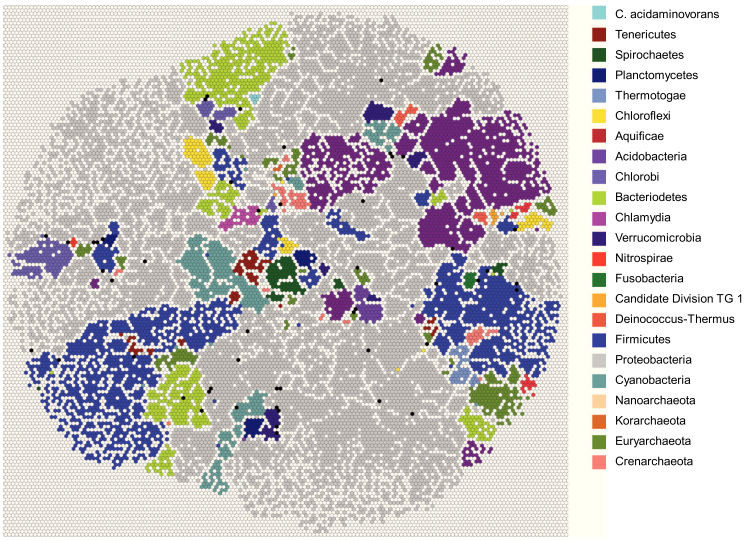


Figure 1 Example of a GSOM showing phylum-level separation. TaxSOM output of a GSOM constructed from all DNA sequences exceeding 485 kb of all *Bacteria* and *Archaea present* in GenBank as of October 2008 (1521 sequences; 3.43 Gb). The figure demonstrates the clustering of sequence fragments of 50 kb with each hexagon representing a single node in the grid. The GSOM was calculated using z-transformed tetranucleotide counts for every fragment. Each color denotes 1 of 23 different phyla, if a node is colored in black it contains fragments of more than one phylum. Nodes displayed in any other color contain only fragments of one particular phylum.

frame to the hit with the highest lineage probability index.

Pfam hits were post-processed with CARMA, an algorithm proposed by Krause *et al.* (2008) that infers taxonomic affiliations from the alignments underlying Pfam Hidden Markov models. Here, we used a rewritten and improved version of the original algorithm.

A weighted consensus of all three tools was used to derive final taxonomic assignments for reads carrying single and contigs carrying multiple genes. The self-written phyloprint C++ library was used to map the taxonomic terms and their NCBI identifiers during the whole analysis.

Algorithm

The SOM is an unsupervised neural network algorithm that implements a non-linear mapping of high-dimensional input data onto a two-dimensional array of weight vectors (Kohonen, 1982, 1990; Kohonen *et al.*, 1996). The process of reducing the data's dimensionality can be thought of as a compression of the input information, whereby the most important topological and metric relationships are preserved. In this sense, SOMs produce an abstraction of the primary data (Kohonen *et al.*, 2001). The topology of the resulting two-dimensional map can be rectangular or hexagonal, and is easy to visualize (Figure 1). Details about input-data variants and a detailed description of the SOM algorithm variants as implemented in TaxSOM are summarized in the Supplementary Methods.

Results

Taxonomic classification of simulated metagenomes

For a close to real-world evaluation, TaxSOM was applied to three published simulated data sets mimicking metagenomes of low, medium and high complexities (simLC, simMC and simHC; see Materials and methods section).

For simLC and simMC, high classification specificities were achieved on both the BLSOM and the GSOM, with almost identical results with either Phrap or Arachne assemblies of at least 8 kb—the length used by Mavromatis *et al.* (2007) in the publication of the simulated metagenome data sets. For GSOM classification of the simLC data set, specificities and sensitivities of 100% were achieved on the superkingdom level and from there on above 97% down to the genus level. For simMC, the classification specificity of the GSOM dropped slightly but stayed above 95% from superkingdom to the genus level, while the sensitivity stayed above 90% (Figure 2a; Supplementary Figure 1). BLSOM classifications yielded almost identical specificities with slightly decreased sensitivities (Supplementary Figure 2a). SimHC was devoid of assemblies exceeding 8 kb and hence was omitted.

When Phrap or Arachne assemblies were used without constraints on sequence size, GSOM

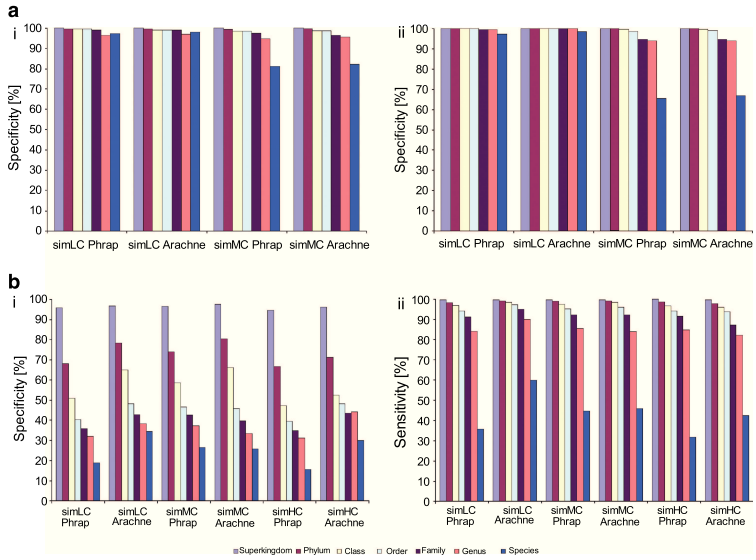


Figure 2 GSOM-based classification specificities of simulated data sets. Taxonomic classification accuracy of TaxSOM for the simulated metagenome data sets mimicking habitats of low (simLC) and medium (simMC) complexities using contigs of 8 kb or larger (**a**) and all contigs (**b**). Plot (i) depicts specificities (%) and plot (ii) sensitivities (%), respectively. From left to right: specificity of classifications of the simLC data sets assembled with PHRAP and Arachne; classifications of the simMC data sets assembled by PHRAP and Arachne. The different taxonomic levels are represented by different colors. All classifications were achieved on a GSOM trained with z-transformed tetranucleotide counts.

classification specificities exceeded 94% (Figure 2b) and those of BLSOM exceeded 96% (Supplementary Figure 2b) on the superkingdom level for all three data sets (simLC, simMC and simHC). Both SOMS were still able to correctly classify >67% of the sequences on the phylum level, while classification accuracy deteriorated notably on deeper taxonomic levels.

Taxonomic classification of data sets from known microorganisms

In order to evaluate the taxonomic classification accuracy of SOMs as a function of DNA fragment lengths, a test data set was constructed from DNA sequences of complete bacterial genome sequences. Parts of the sequences were used to construct di-, tri- and tetranucleotide-based SOMs and the remainder was split into fragments of different lengths and subsequently classified using the SOMs (see Materials and methods section).

Classification specificity improved with increasing motif and fragment lengths (Supplementary Table 1). It was mostly above 80% for sequences of at least 5 kb and even above 90% for longer fragments on low-resolution taxonomic levels. Below 5 kb classification specificities quickly dropped to values of mostly below 50%, especially for high-resolution genus and species assignments. One interesting observation was that for fragments of 5 kb or more, z-scores provided better assignments while below 5 kb, raw scores provided more accurate results. Also, GSOMs performed better than BLSOMs. Generally speaking, high-resolution assignments required longer sequences (that is, higher information content) than broad-level assignments. For instance, in order to have a 70% accuracy with dinucleotide-based GSOMs, sequences <0.5 kb were sufficient on the superkingdom level, 2.5 kb on the phylum level, 5 kb on the class level, 25 kb on the family level and >50 kb on the genus level. Similar patterns were observed with longer motif lengths, although longer motifs increased classification accuracy. For example, based on tetranucleotides, 70% classification accuracy on the genus level was possible with <10 kb (Supplementary Table 1).

Taxonomic classification of real-world metagenome data sets

TaxSOM's ability to classify real-world metagenome data was assessed with five North Sea metagenome data sets comprising a total of nine full 454 FLX Ti pyrosequencing runs amounting to almost 8 million reads (Table 1). These could be assembled into 340 143 contigs, of which those of 2.5 kb or more were used for classification. The sample taken in February was highest in biodiversity and yielded only 227 contigs of sufficient length (0.8 Mb). In contrast, the samples taken in March and April had lower biodiversities since they covered a *Bacteroidetes*

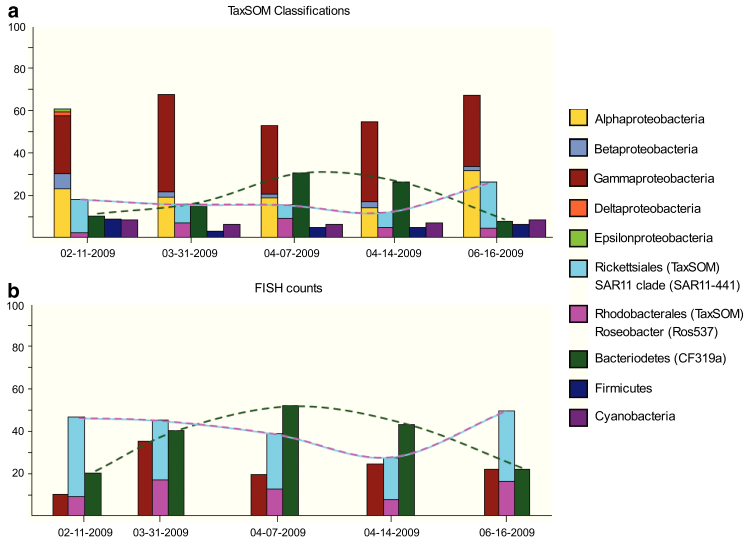


Figure 3 Biodiversity assessments of the North Sea metagenomes over time. (**a**) Taxonomic classification of assemblies exceeding 2.5 kb with TaxSOM. (**b**) Relative CARD-FISH counts of corresponding water samples. Dotted lines indicate congruence in the abundances of Bacteroidetes (green) and the majority of *Alphaproteobacteria* consisting of the orders *Rickettsiales* (cyan) and *Rhodobacterales* (magenta) as assessed by both methods. They do not indicate a smooth transition of the respective abundances, because the community composition fluctuated considerably in between sample time points (data not shown).

bloom, and thus could be assembled into 2321–3229 contigs of sufficient lengths (9.8–16.2 Mb), while the last sample taken in June yielded 1137 such contigs (5 Mb).

In order to assess the plausibility of TaxSOM's taxonomic classifications, we compared the classifications with corresponding CARD-FISH counts of water samples for all five data sets (Figure 3). In the February post-winter situation, the water was low in temperature (~4 °C) and cell densities (4 E6 cells ml^{-1}). Only in this diverse sample TaxSOM detected *Deltaproteobacteria* and *Epsilonproteobacteria*, which were likely dispersed from the sediment by winter storm perturbations.

The spring situation from end of March to mid-April was characterized by a slight increase in water temperature (~6 °C) and cell densities (~1 E6 cells ml^{-1}). TaxSOM and the CARD-FISH data both detected a spring bloom in *Bacteroidetes* that reached a maximum in mid-April and was accompanied by a decrease in *Alphaproteobacteria* from the SAR11 and *Roseobacter* clades. Hence, much of the original biodiversity patterns were retained in the sequence assemblies. Of course, absolute numbers differed. For example, after the bloom maximum in mid-April, equal levels of *Gammaproteobacteria* (25%) and *Alphaproteobacteria* (27%, SAR11 and *Roseobacter* combined), and much higher abundances of *Bacteroidetes* (43%) were detected *in situ* with CARD-FISH, while in the assemblies TaxSOM detected more *Gammaproteobacteria* (36%) than *Alphaproteobacteria* (13%) and *Bacteroidetes* (26%).

In addition to CARD-FISH, we compared TaxSOM's classifications with protein-based classifications for the data sets of mid-April (Supplementary Figure 3). Both allow a complete taxonomic breakdown from the superkingdom to the species level. Again, while absolute numbers are different from those of FISH *in situ* measurements, the overall biodiversity pattern was retained. In comparison with PBC, only TaxSOM was able to resolve the high abundances of *Bacteroidetes* (TaxSOM: 28%; PBC: 13%). Both tools were able to resolve the key players down to the genus level. Most of the *Bacteroidetes* were resolved as *Polaribacter*-like *Flavobacteria* by both tools. Similarly, a large proportion of the *Alphaproteobacteria* was resolved as SAR11 and *Roseobacter* clade species, as indicated by hits to *Pelagibacter* and *Roseobacter* on the genus level. This is in line with the CARD-FISH results as well as with reported high abundance of SAR11 species in the oceans by previous metagenome studies (Temperton *et al.*, 2009).

Complete taxonomic breakdowns of all five metagenome data sets are included in the Supplementary Material of this study.

It is noteworthy that as a signature-based method, TaxSOM could classify the contigs without suitable BLAST and HMMer hits that could not be classified on the level of proteins. Especially on deeper taxonomic levels, sequences could oftentimes not

be classify based on protein information but could be classified by TaxSOM, which thus provided a much more detailed taxonomic breakdown.

Discussion

In this study, we demonstrate that DNA composition-based SOMs as implemented in TaxSOM are a valuable and useful tool for the taxonomic classification of microbial metagenomes and their subsequent ecological interpretation. Most suitable in this respect are NGS-based deeply sequenced metagenomes of habitats with a low to medium biodiversity, as for example in pelagic ocean waters.

Simulated metagenomes
When applied to simulated metagenome data sets, TaxSOM achieved high classification specificities down to the genus level for the data sets mimicking low- and medium-diversity habitats with fragments of at least 8 kb. These results were obtained even though the corresponding SOMs were constructed from all available fully sequenced prokaryote genomes, and thus comprised a wealth of nodes representing species lacking from the simulated data sets, leading to a high statistical chance of misclassification. This implies that with real-world data from habitats of comparable complexities, respective fragments can be classified with specificities that are sufficient to deduce biologically meaningful results down to the family or even genus level. As in most real-world applications *a priori* knowledge about the studied habitat is available, more specific SOMs can be constructed from dedicated training sequences, which will further improve classification specificity. Using the simulated data sets without constraining fragment lengths lead to a notable decrease in classification specificities. One reason for this is of course that without length restrictions, large quantities of very short sequences were included whose information content is insufficient for accurate classification. Interestingly, this effect was almost independent of the complexity of the simulated data set, suggesting that at least these data sets were not saturating the resolution of the SOM, that is, the complexity of the analyzed data was not limiting the analysis. An additional reason for the drop in classification specificities, as stated by Mavromatis *et al.*, (2007), is that a high proportion of chimeras among shorter contigs result in low quality classifications. If the number of such misassemblies can be reduced, the minimum required sequence length will drop as well (Chan *et al.*, 2008b). Still, even with inclusion of the short fragments TaxSOM provided respectable results in all simulated data sets at least down to the phylum level, which might be the current limit for reasonable biological conclusions based on mostly short and unassembled sequences (Figure 2; Supplementary Figures 1 and 2).

Data sets from known microorganisms
The results from the artificial data sets of fully sequenced microorganisms show that classification specificity is a function of information content, and hence increases with motif and with sequence length. Longer oligonucleotides provide better specificities than shorter oligonucleotides, and longer sequences can be classified more accurately than shorter ones. Good classification specificities can be obtained for sequences down to 5 kb; below that, information content starts to become limiting (Supplementary Table 1). This is also supported by the fact that below 5 kb SOMs constructed from raw oligonucleotide counts outperformed those constructed from z-transformed counts, while it was the opposite above 5 kb. The z-transformation statistically corrects counts of oligonucleotides of a given length for asymmetries introduced by skews in shorter oligonucleotide frequencies. For example, it is expected that within high GC genomes higher frequencies of GC-rich tetranucleotides (for example, GGCC) are observed than an AT-rich genomes, and thus high frequencies of GGCC in an GC-rich genome convey less information as when they occur in an AT-rich genome. However, the z-transformation compensating this is itself based on a statistical assessment, and hence also limited by the sequence's information content. As the latter deteriorates from about 5 kb on, z-transformation can only enhance results for sequences with sufficient information content and even introduces additional noise when the sequences get too short for proper statistics. Nonetheless, classification accuracies for sequences below 5 kb are still sufficient to conduct NGS-based statistical ecological habitat studies. Here, the ability to discriminate a biological signal from the data's noise is more important than an almost perfect classification, such as when monitoring overall community composition changes or linking abundances of functional genes to taxonomic groups.

Real-world metagenomes
As our results with pyrosequenced bulk DNA show, such studies are possible with sequences of 2.5 kb at least down to the class if not to the order level, especially with suitable habitat-specific SOMs. Although classifications were not perfect with respect to providing a high-resolution quantitative taxonomic breakdown of the analyzed samples, they provide a good description of overall biodiversity and abundances of a given habitat (Supplementary Figure 3) and allow detection of major community composition changes (Figure 3). These data can serve as a guideline for the selection of specific CARD-FISH probes for more detailed biodiversity studies, and furthermore allows mining the taxonomically classified sequences (taxobins) for functions. Such a linkage of taxonomy and function will allow us to gain insights into the ecological

functioning of habitats and even to select frequent but as yet unknown genes within dedicated taxa as targets for further studies.

It is our experience from more than a dozen direct pyrosequencing experiments on moderately diverse coastal and deep sea ocean waters (data not shown) that well-run 454 FLX Ti picotiter plates can yield more than a million reads comprising up to 400 Mb of raw sequence that typically can be assembled into 30–70 Mb of non-redundant DNA, equivalent to 7–16 bacterial genomes. In all cases, the longest assemblies were well within the range of typical fosmids (that is, up to 35 kb), and larger proportions of the assemblies were above 2.5 kb and thus suitable for SOM analysis (Table 1).

Biodiversity information from direct DNA sequencing cannot rival *in situ* measurements like FISH in terms of quantitativeness because of inherent biases, such as lineage-specific DNA extraction efficacies, sequence-dependent differences during the bead-mediated amplification in the 454 library creation step, skews introduced by the assembly diminishing quantities of the most abundant species, and taxa without suitable reference sequences for taxonomic classification, like those without any representation in public sequence databases. FISH on the other hand has to cope with its own inherent limits, like issues with permeabilization, target accessibility or probe sensitivity and specificity. Hence, both methods shed a slightly different light on biodiversity. It is therefore understandable that the biodiversity data obtained by FISH and by direct sequencing of bulk DNA show differences, although they are in broad agreement with respect to major community composition shifts. It is noteworthy that the TaxSOM assignments were well supported by PBC tools. This indicates that the TaxSOM assignments reflect a realistic assessment of the biodiversity within the sampled sequences, which does not necessarily reflect the situation *in situ* in a perfect manner. FISH does provide only information for the applied probes, whereas *in silico* taxonomic classifications of directly sequenced DNA do not require *a priori* assumptions about the community composition, provide a deeper taxonomic resolution in shorter time and enable formation of taxobins that can be mined for gene functions in order to address ecological questions.

TaxSOM website

In order to make such applications accessible for a broader audience in microbial molecular ecology, we implemented TaxSOM as a freely accessible website that allows the usage of GSOMs and BLSOMs for taxonomic classification of microbial DNA sequences. TaxSOM provides either pre-computed SOMs for general taxonomic classification purposes, or the option to compute custom-tailored SOMs. For the latter, TaxSOM provides the ability to upload sequences for SOM construction (for example, with habitat-specific sequences) as well as a dynamic taxonomy tree selection tool that allows for an easy visual as well as textual selection of all sequences of the NCBI nt database with sufficient length. A rich set of features is available for controlling the behavior of SOMs, and the resulting SOMs can be inspected visually. For experts, we provide a rich set of parameters for controlling the SOMs behavior. Unique to TaxSOM is the capability of pre-processing frequencies using a maximal-order Markov model as input data, which improves classification accuracy for sequences exceeding 5 kb. After a SOM is constructed, sequences can be uploaded for classification, whereby a SOM persists and can be used for the classification of multiple data sets. Classification results can be inspected either visually (Figure 1) or downloaded as tables in text files for further use. This will enable a broader audience to use taxonomic classifications in microbial community studies. The TaxSOM web service is available at http://www.megx.net/toolbox/taxsom.

Conclusions

One advantage of SOMs is that taxonomic classification once a SOM is trained takes only minutes, even for large amounts of sequences, while gene-based classification tools rely on time-consuming and computationally intensive BLAST or HMMER searches, and FISH requires labor-intensive laboratory work. For example, the current TaxSOM implementation can classify 100 000 sequences on a SOM of 10 000 nodes within 20 min on moderate hardware (single 2.2 MHz Opteron core). Similarly, a million sequences can be classified within a couple of hours conveniently over night (see Supplementary Tables 4, 5 and 6 for more elaborate data on classification speed). This makes SOM-based taxonomic classifications ideal for processing vast amounts of sequences as they are produced by current NGS platforms.

With promising new sequencing technologies on the horizon that will not need amplification and will deliver more and longer reads at lower prices, like ZMW-based sequencing by Pacific Biosciences (Menlo Park, CA, USA; Eid *et al.*, 2009) or various variants of nanopore sequencing as developed by Oxford Nanopore Technologies (Kidlington, Oxfordshire, UK; Clarke *et al.*, 2009), IBM Deutschland Research & Development GmbH (Böblingen, Germany)/Roche (DNA transistor) and others, there is a need for high-throughput tools to convert the wealth of sequence data into knowledge. The recently introduced Pacific Biosciences single molecule sequencing platform has an average read length of >1 kb and a maximum read length of 5 kb. Hence, technologies that focus on short reads will be mostly obsolete in the not too distant future. As a consequence, microbial biodiversity studies soon will target full-length 16S ribosomal RNA sequences instead of only small hypervariable regions, and

metagenomic studies will produce longer assemblies that can be taxonomically classified with high accuracy. Tools like TaxSOM will enable the fast classification of large proportions of metagenomes into taxobins, and thus provide a link between biodiversity and function.

Even with current 454 FLX Titanium pyrosequencing, good results can be expected for SOM-based taxonomic classifications, in particular for habitats with limited diversity, few dominating species or with species that discriminate well in terms of their genomic signatures. For complex habitats leading to metagenomes without longer assemblies, DNA composition-based methods should be combined with if not substituted by PBC methods. These, however, are restricted to sequences harboring well-characterized genes or domains and thus can classify fewer sequences.

Until long read technologies are available, we suggest clustering metagenomes into taxobins by a combination of nucleotide and protein-based taxonomic classification tools. This enables the application of large-scale NGS DNA sequencing as a screening tool for biodiversity and paves the way for insights into the functional ecology of complex microbial communities. For habitats with low-to-medium biodiversity, sufficiently reliable classifications can be achieved down to the genus level, but the amount of sequence that is obtained in praxis with current techniques will be often too small for a sound statistical analysis of gene functions on this level. However, functional analyses of our real-world data have shown that such studies can be done down to the class and for abundant taxa even down to the order level with two to four full picotiter plates of pyrosequencing per sample (data not shown). We anticipate that progress in sequencing with respect to read length and throughput will soon eliminate this bottleneck and thus will enable to study microbial communities in a holistic fashion on a much finer level.

Acknowledgements

We thank Tobin J Hammer for fruitful discussions and proof reading of the paper. This study was supported by the Max Planck society and the MIMAS project (project no. 03F0480A) funded by the German Federal Ministry of Education and Research (BMBF).

References

Abe T, Sugawara H, Kinouchi M, Kanaya S, Ikemura T. (2005). Novel phylogenetic studies of genomic sequence fragments derived from uncultured microbe mixtures in environmental and clinical samples. *DNA Res* **12**: 281–290.

Abe T, Kanaya S, Kinouchi M, Ichiba Y, Kozuki T, Ikemura T. (2003). Informatics for unveiling hidden genome signatures. *Genome Res* **13**: 693–702.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. (1990). Basic local alignment search tool. *J Mol Biol* **215**: 403–410.

Amann RI, Ludwig W, Schleifer KH. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev* **59**: 143–169.

Brady A, Salzberg SL. (2009). Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. *Nat Methods* **6**: 673 .

Burge C, Campbell AM, Karlin S. (1992). Over- and under-representation of short oligonucleotides in DNA sequences. *Proc Natl Acad Sci USA* **89**: 1358–1362.

Chan C-KK, Hsu AL, Tang S-L, Halgamuge SK. (2008a). Using growing self-organising maps to improve the binning process in environmental whole-genome shotgun sequencing. *J Biomed Biotechnol* **2008**, doi:10.1155/2008/513701.

Chan CK, Hsu AL, Halgamuge SK, Tang SL. (2008b). Binning sequences using very sparse labels within a metagenome. *BMC Bioinformat* **9**: 215.

Clarke J, Wu HC, Jayasinghe L, Patel A, Reid S, Bayley H. (2009). Continuous base identification for single-molecule nanopore DNA sequencing. *Nat Nanotechnol* **4**: 265–270.

Deschavanne PJ, Giron A, Vilain J, Fagot G, Fertil B. (1999). Genomic signature: characterization and classification of species assessed by chaos game representation of sequences. *Mol Biol Evol* **16**: 1391–1399.

Diaz NN, Krause L, Goesmann A, Niehaus K, Nattkemper TW. (2009). TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *BMC Bioinformat* **10**: 56.

Dick GJ, Andersson AF, Baker BJ, Simmons SL, Thomas BC, Yelton AP *et al.* (2009). Community-wide analysis of microbial genome sequence signatures. *Genome Biol* **10**: R85.

Eddy SR. (1996). Hidden Markov models. *Curr Opin Struct Biol* **6**: 361–365.

Eddy SR. (1998). Profile Hidden Markov Models. *Bioinformatics* **14**: 755–763.

Eid J, Fehr A, Gray J, Luong K, Lyle J, Otto G *et al.* (2009). Real-time DNA sequencing from single polymerase molecules. *Science* **323**: 133–138.

Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR *et al.* (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science* **269**: 496–512.

Gupta PK. (2008). Single-molecule DNA sequencing technologies for future genomics research. *Trends Biotechnol* **26**: 602–611.

Hanekamp K, Bohnebeck U, Beszteri B, Valentin K. (2007). PhyloGena—a user-friendly system for automated phylogenetic annotation of unknown sequences. *Bioinformatics* **23**: 793–801.

Huber JA, Mark Welch DB, Morrison HG, Huse SM, Neal PR, Butterfield DA *et al.* (2007). Microbial population structures in the deep marine biosphere. *Science* **318**: 97–100.

Huse SM, Dethlefsen L, Huber JA, Welch DM, Relman DA, Sogin ML. (2008). Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet* **4**: e1000255.

Huson DH, Auch AF, Qi J, Schuster SC. (2007). MEGAN analysis of metagenomic data. *Genome Res* **17**: 377–386.

Jaffe DB, Butler J, Gnerre S, Mauceli E, Lindblad-Toh K, Mesirov JP *et al.* (2003). Whole-genome sequence assembly for mammalian genomes: Arachne 2. *Genome Res* **13**: 91–96.

Karlin S, Burge C. (1995). Dinucleotide relative abundance extremes: a genomic signature. *Trends Genet* **11**: 283–290.

Karlin S. (1998). Global dinucleotide signatures and analysis of genomic heterogeneity. *Curr Opin Microbiol* **1**: 598–610.

Karlin S, Campbell AM, Mrazek J. (1998). Comparative DNA analysis across diverse genomes. *Annu Rev Genet* **32**: 185–225.

Karlin S, Ladunga I. (1994). Comparisons of eukaryotic genomic sequences. *Proc Natl Acad Sci USA* **91**: 12832–12836.

Karlin S, Ladunga I, Blaisdell BE. (1994). Heterogeneity of genomes: measures and values. *Proc Natl Acad Sci USA* **91**: 12837–12841.

Kohonen T. (1982). Self-organized formation of topologically correct feature maps. *Biol Cybernet* **43**: 59–69.

Kohonen T. (1990). Self-organization maps. *Proc IEEE* **78**: 1464–1480.

Kohonen T, Kohonen T, Schroeder MR, Huang TS, Maps SO. (2001). Springer-Verlag New York Inc.: Secaucus, NJ.

Kohonen T, Oja E, Simula O, Visa A, Kangas J. (1996). Engineering applications of the self-organizing map. *Proc IEEE* **84**: 1358–1384.

Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W *et al.* (2010). Megx net: integrated database resource for marine ecological genomics. *Nucleic Acids Res* **38**: D391–D395.

Krause L, Diaz NN, Goesmann A, Kelley S, Nattkemper TW, Rohwer F *et al.* (2008). Phylogenetic classification of short environmental DNA fragments. *Nucleic Acids Res* **36**: 2230 .

Martin C, Diaz NN, Ontrup J, Nattkemper TW. (2008). Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. *Bioinformatics* **24**: 1568–1574.

Mavromatis K, Ivanova N, Barry K, Shapiro H, Goltsman E, McHardy AC *et al.* (2007). Use of simulated data sets to evaluate the fidelity of metagenomic processing methods. *Nat Methods* **4**: 495–500.

McHardy AC, Martin HG, Tsirigos A, Hugenholtz P, Rigoutsos I. (2007). Accurate phylogenetic classification of variable-length DNA fragments. *Nat Methods S* **4**: 63–72.

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

Ochman H. (2007). Single-cell genomics. *Environ Microbiol* **9**: 7.

Pernthaler A, Pernthaler J, Amann R. (2002). Fluorescence in situ hybridization and catalyzed reporter deposition for the identification of marine bacteria. *Appl Environ Microbiol* **68**: 3094–3101.

Peterson J, Garges S, Giovanni M, McInnes P, Wang L, Schloss JA *et al.* (2009). The NIH Human Microbiome Project. *Genome Res* **19**: 2317–2323.

Podell S, Gaasterland T. (2007). DarkHorse: a method for genome-wide prediction of horizontal gene transfer. *Genome Biol S* **8**: R16.

Pride DT, Meinersmann RJ, Wassenaar TM, Blaser MJ. (2003). Evolutionary implications of microbial genome tetranucleotide frequency biases. *Genome Res* **13**: 145–158.

Reva ON, Tümmler B. (2004). Global features of sequences of bacterial chromosomes, plasmids and phages revealed by analysis of oligonucleotide usage patterns. *BMC Bioinformat* **5**: 90.

Rocha EP, Viari A, Danchin A. (1998). Oligonucleotide bias in Bacillus subtilis: general trends and taxonomic comparisons. *Nucleic Acids Res* **26**: 2971–2980.

Sandberg R, Winberg G, Branden CI, Kaske A, Ernberg I, Coster J. (2001). Capturing whole-genome characteristics in short sequences using a naive Bayesian classifier. *Genome Res* **11**: 1404–1409.

Schloss PD, Handelsman J. (2003). Biotechnological prospects from metagenomics. *Curr Opin Biotechnol* **14**: 303–310.

Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M. (2007). CAMERA: a community resource for metagenomics. *PLoS Biol* **5**: e75.

Sogin ML, Morrison HG, Huber JA, Mark Welch D, Huse SM, Neal PR *et al.* (2006). Microbial diversity in the deep sea and the underexplored 'rare biosphere'. *Proc Natl Acad Sci USA* **103**: 12115–12120.

Sonnhammer EL, Eddy SR, Durbin R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. *Proteins* **28**: 405–420.

Sonnhammer EL, Eddy SR, Birney E, Bateman A, Durbin R. (1998). Pfam: multiple sequence alignments and HMM-profiles of protein domains. *Nucleic Acids Res* **26**: 320–322.

Teeling H, Meyerdierks A, Bauer M, Amann R, Glöckner FO. (2004). Application of tetranucleotide frequencies for the assignment of genomic fragments. *Environ Microbiol* **6**: 938–947.

Temperton B, Field D, Oliver A, Tiwari B, Muhling M, Joint I *et al.* (2009). Bias in assessments of marine microbial biodiversity in fosmid libraries as evaluated by pyrosequencing. *ISME J* **3**: 792–796.

Tringe SG, von Mering C, Kobayashi A, Salamov AA, Chen K, Chang HW *et al.* (2005). Comparative metagenomics of microbial communities. *Science S* **308**: 554–557.

Tyson GW, Chapman J, Hugenholtz P, Allen EE, Ram RJ, Richardson PM *et al.* (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature* **428**: 37–43.

Waldmann J. (2008). Phyloprint—Entwicklung und Anwendung eines Frameworks zur taxonomischen Klassifikation. *Westfälische Wilhelms-Universität Münster, Department of Mathematics and Computer Science, Diploma Thesis*, http://cs.uni-muenster.de/Professoren/Lippe/diplomarbeiten/html/Waldmann/.

Woyke T, Teeling H, Ivanova NN, Huntemann M, Richter M, Gloeckner FO, *et al.* (2006). Symbiosis insights through metagenomic analysis of a microbial consortium. *Nature* **443**: 950–955.

Wu D, Hugenholtz P, Mavromatis K, Pukall R, Dalin E, Ivanova NN *et al.* (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature* **462**: 1056–1060.

Zhou J, Bruns MA, Tiedje JM. (1996). DNA recovery from soils of diverse composition. *Appl Environ Microbiol S* **62**: 316–322.

Supplementary Information accompanies the paper on The ISME Journal website (http://www.nature.com/ismej)

2.2 Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom

Authors:

Hanno Teeling[1□], Bernhard M. Fuchs[1□], D□rte Becher[2,5], Christine Klockow[1,3], Antje Gardebrecht[6], Christin M. Bennke[1], Mariette Kassabgy[1], Sixing Huang[1], Alexander J. Mann[1,3], Jost Waldmann[1,2,3], Marc Weber[1,3], Anna Klindworth[1,3], Andreas Otto[5], Jana Lange[2], J□rg Bernhardt[5,7], Christine Reinsch[2], Michael Hecker[2,5], J□rg Peplies[8], Frank D. Bockelmann[9], Ulrich Callies[9], Gunnar Gerdts[4], Antje Wichels[4], Karen H. Wiltshire[4], Frank Oliver Gl□ckner[1,3], Thomas Schweder[2,6]*, and Rudolf Amann[1]*

[1] Max Planck Institute for Marine Microbiology, Celsiusstr. 1, 28359 Bremen, Germany
[2] Institute of Marine Biotechnology e.V., Walther-Rathenau-Str. 49a, 17489 Greifswald, Germany
[3] Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen, Germany
[4] Alfred Wegener Institute for Polar and Marine Research, Biologische Anstalt Helgoland, 27483 Helgoland, Germany
[5] Institute for Microbiology, Ernst-Moritz-Arndt University, Friedrich-Ludwig- Jahn-Str. 15, 17487 Greifswald, Germany
[6] Pharmaceutical Biotechnology, Ernst-Moritz-Arndt University, Felix- Hausdorff-Str. 3, 17487 Greifswald, Germany
[7] DECODON GmbH, Walther-Rathenau-Str. 49a, 17489 Greifswald, Germany
[8] Ribocon GmbH, Fahrenheitstr. 1, 28359 Bremen, Germany
[9] HZG Research Center, Max-Planck Str. 1, 21502 Geesthacht, Germany

□ These authors contributed equally to this work

□Corresponding authors: Thomas Schweder and Rudolf Amann

Publication status:

Published in □□□□□□

My contribution:

In this project, I was responsible for the processing of the assembled metagenomic sequences, including taxonomic classification and functional annotation. I implemented Kirsten, CARMA and DarkHorse that were integrated parts of the taxonomic classification pipeline. I generated the CAZyme profiles for the metagenomes and calculated the expression levels based on the metatranscriptomic and metaproteomic data. Finally, I was involved in the discussion and interpretation of the results.

13. R. G. Thorne, C. Nicholson, *Proc. Natl. Acad. Sci. U.S.A.* **103**, 5567 (2006).
14. G. E. Hardingham, H. Bading, *Microsc. Res. Tech.* **46**, 348 (1999).
15. A. Rao, *Nat. Immunol.* **10**, 3 (2009).
16. A. L. Shifrin, A. Auricchio, Q. C. Yu, J. Wilson, S. E. Raper, *Gene Ther.* **8**, 1480 (2001).
17. M. Tewari *et al.*, *Cell* **81**, 801 (1995).
18. Y. Gavrieli, Y. Sherman, S. A. Ben-Sasson, *J. Cell Biol.* **119**, 493 (1992).
19. H. Huang, S. Delikanli, H. Zeng, D. M. Ferkey, A. Pralle, *Nat. Nanotechnol.* **5**, 602 (2010).
20. J. L. Farrant, *Biochim. Biophys. Acta* **13**, 569 (1954).
21. B. Sana, E. Johnson, K. Sheah, C. L. Poh, S. Lim, *Biointerphases* **5**, FA48 (2010).
22. K. Ziv *et al.*, *NMR Biomed.* **23**, 523 (2010).
23. B. Iordanova, C. S. Robison, E. T. Ahrens, *J. Biol. Inorg. Chem.* **15**, 957 (2010).
24. D. Halperin *et al.*, in *Proceedings of the 2008 IEEE Symposium on Security and Privacy*, Oakland, CA, 18 to 21 May 2008, pp. 129–142.
25. S. A. Hanna, paper presented at the Third International Symposium on Medical Information and Communication Technology, Montreal, Canada, 24 to 27 February 2009.
26. A. C. Nathwani *et al.*, *N. Engl. J. Med.* **365**, 2357 (2011).
27. C. A. Butts *et al.*, *Biochemistry* **47**, 12729 (2008).

Acknowledgments: We thank Friedman laboratory members for helpful discussions, S. Korres for assistance with manuscript preparation, S. Tavazoie for helpful discussions, R. Toledo-Crow and S. Abeytunge for assistance with RF technology, and the staff of the Rockefeller University Electron Microscopy Resource center for their technical support in imaging. This project was supported by funding from the JPB Foundation and grant no. R01 GM095654 from the NIH. The authors have filed a patent related to this work.

Supplementary Materials
www.sciencemag.org/cgi/content/full/336/6081/604/DC1
Materials and Methods
Supplementary Text
Figs. S1 to 10
References (*28*–*35*)

17 November 2011; accepted 23 March 2012
10.1126/science.1216753

Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom

Hanno Teeling,[1]* Bernhard M. Fuchs,[1]* Dörte Becher,[2,5] Christine Klockow,[1,3] Antje Gardebrecht,[6] Christin M. Bennke,[1] Mariette Kassabgy,[1] Sixing Huang,[1] Alexander J. Mann,[1,3] Jost Waldmann,[1,2,3] Marc Weber,[1,3] Anna Klindworth,[1,3] Andreas Otto,[5] Jana Lange,[2] Jörg Bernhardt,[5,7] Christine Reinsch,[2] Michael Hecker,[2,5] Jörg Peplies,[8] Frank D. Bockelmann,[9] Ulrich Callies,[9] Gunnar Gerdts,[4] Antje Wichels,[4] Karen H. Wiltshire,[4] Frank Oliver Glöckner,[1,3] Thomas Schweder,[2,6]† Rudolf Amann[1]†

Phytoplankton blooms characterize temperate ocean margin zones in spring. We investigated the bacterioplankton response to a diatom bloom in the North Sea and observed a dynamic succession of populations at genus-level resolution. Taxonomically distinct expressions of carbohydrate-active enzymes (transporters; in particular, TonB-dependent transporters) and phosphate acquisition strategies were found, indicating that distinct populations of *Bacteroidetes*, *Gammaproteobacteria*, and *Alphaproteobacteria* are specialized for successive decomposition of algal-derived organic matter. Our results suggest that algal substrate availability provided a series of ecological niches in which specialized populations could bloom. This reveals how planktonic species, despite their seemingly homogeneous habitat, can evade extinction by direct competition.

Annually recurring spring phytoplankton blooms with high net primary production (NPP) characterize eutrophic upwelling zones and coastal oceans in higher latitudes. Coastal zones with water depths <200 m constitute ~7% of the global ocean surface (*1*), yet they are responsible for ~19% of the oceanic NPP (*2*) and globally account for 80% of organic matter burial and 90% of sedimentary mineralization (*1*). Heterotrophic members of the picoplankton—mostly *Bacteria*—reprocess about half of the oceanic NPP in the so-called "microbial loop" (*3*). The bulk of this bacterioplankton biomass is free-living, but up to 20% is attached to algae or particles (*4*).

The bacterial response to coastal phytoplankton blooms has been almost exclusively studied in microcosm/mesocosm experiments (*5*–*8*) or with limited resolution in time and biodiversity in situ (*9*–*11*). We observed bacterial populations during and after a phytoplankton bloom in spring 2009 at the island of Helgoland in the German Bight (54°11′03″N, 7°54′00″E; fig. S1A) with a high taxonomic and functional resolution. We sampled 500 liters of subsurface seawater twice a week during 2009. Samples were filtered into fractions dominated by free-living bacteria (3 to 0.2 μm in size) and algae/particle-associated bacteria (10 to 3 μm in size) (fig. S2). Algal composition was determined microscopically (fig. S3 and table S1), and microbial composition was identified via catalyzed reporter deposition fluorescence in situ hybridization (CARD-FISH, tables S2 and S3). At selected sampling times during and after the bloom, the data were complemented by comparative analysis of 16*S* ribosomal RNA (rRNA) gene amplicons (pyrotags, table S4) and by functional data from extensive metagenome and metaproteome analyses (table S5). In addition, physical and chemical parameters were measured daily, including temperature, turbidity, salinity, and concentrations of phosphate, nitrate, nitrite, ammonium, silicate, and chlorophyll a (table S6).

Pre-bloom bacteria (Fig. 1A) were dominated by *Alphaproteobacteria* (41 to 67%), composed roughly of two-thirds SAR11 clade and one-third *Roseobacter* clade (Fig. 1B and fig. S4B). SAR11 consisted almost exclusively of subgroup Ia (*Candidatus Pelagibacter ubique*) (table S4). This composition changed as the spring phytoplankton bloom commenced (*12*). In early April (3 to 9 April 2009), *Bacteroidetes* abundances increased fivefold within 1 week (from 1.5×10^5 to 7.7×10^5 cells/ml), whereas *Alphaproteobacteria* (from 2.1×10^5 to 5.0×10^5 cells/ml) and *Gammaproteobacteria* (from 0.8×10^5 to 1.8×10^5 cells/ml) abundances only approximately doubled. The *Bacteroidetes* consisted mostly of *Flavobacteria* (89 to 98%) (table S4), with a succession of *Ulvibacter* spp., followed by *Formosa*-related and *Polaribacter* species as the most prominent groups (Fig. 1C and fig. S4C). *Gammaproteobacteria* reacted later to algal decay, but with a more dense succession of peaking clades, with highest abundances in *Reinekea* spp. and SAR92 (Fig. 1D and fig. S4D). *Reinekea* spp. grew within 1 week from 1.6×10^3 cells/ml to above 1.6×10^5 cells/ml (estimated doubling time, 25 hours) and subsequently almost vanished within 2 weeks. *Roseobacter* clade members also showed a succession, with the NAC11-7 lineage dominating the early bacterioplankton bloom and the *Roseobacter* clade–affiliated (RCA) lineage dominating the late bloom (table S4).

Metagenomes were partitioned into taxonomically coherent bins (taxobins, fig. S5A) and then used for identification, annotation, and semiquantitative analyses of the metaproteome data (*12*). This allowed the investigation of shifts in gene content and expression within dominating bacterial populations (table S7).

[1]Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, 28359 Bremen, Germany. [2]Institute of Marine Biotechnology, Walther-Rathenau-Strasse 49a, 17489 Greifswald, Germany. [3]Jacobs University Bremen, Campus Ring 1, 28759 Bremen, Germany. [4]Alfred Wegener Institute for Polar and Marine Research, Biologische Anstalt Helgoland, 27483 Helgoland, Germany. [5]Institute for Microbiology, Ernst-Moritz-Arndt University, Friedrich-Ludwig-Jahn-Strasse 15, 17487 Greifswald, Germany. [6]Pharmaceutical Biotechnology, Ernst-Moritz-Arndt University, Felix-Hausdorff-Strasse 3, 17487 Greifswald, Germany. [7]DECODON, Walther-Rathenau-Strasse 49a, 17489 Greifswald, Germany. [8]Ribocon, Fahrenheitstrasse 1, 28359 Bremen, Germany. [9]HZG Research Center, Max-Planck Strasse 1, 21502 Geesthacht, Germany.

*These authors contributed equally to this work.
†To whom correspondence should be addressed. E-mail: schweder@uni-greifswald.de (T.S.); ramann@mpi-bremen.de (R.A.)

A pronounced peak in the abundance of carbohydrate-active enzymes [CAZymes (*13*)] accompanied the bacterial succession (fig. S5B). CAZyme frequencies and expressions were taxonomically distinct (Figs. 2 and 3). For instance, *Flavobacteria* and *Gammaproteobacteria* dominated the abundant glycoside hydrolase family 16 (GH16). Most corresponding genes were annotated as laminarinases for decomposing the algal glucan laminarin. Likewise, expressed GH30-family proteins that include β-D-fucosidases mapped exclusively to *Flavobacteria*. *Flavobacteria* also dominated GH29/GH95-family genes containing α-L-fucosidases, as well as L-fucose permease genes. Fucose is a major constituent of diatom exopolysaccharides (*14, 15*). *Flavobacteria* were also dominating GH92-family glycoside hydrolases encoding mainly alpha-mannosidase, whereas *Gammaproteobacteria* dominated the glycoside hydrolase family 81. Likewise, *Gammaproteobacteria* (SAR92 clade) and *Flavobacteria* dominated expression within the GH3 family.

Many algal polysaccharides are sulfated (such as carragenans, agarans, ulvans, and fucans), and hence sulfatases are required for their complete degradation. Sulfatase gene frequencies peaked together with the CAZymes at 7 April and showed a mixed taxonomic composition, but the maximum in sulfatase expression occurred later in the bloom (Fig. 3) and was dominated by *Flavobacteria*. Expressed sulfatases were found in the *Polaribacter* taxobin, which corroborates recent reports of high numbers of sulfatases in *Polaribacter* (*16*). In contrast, glycoside hydrolases for decomposing nonsulfated laminarin (GH16, GH55, and GH117) had their expression maxima earlier during the initial algal die-off phase.

Glycolytic exoenzymes initiate bacterial utilization of complex algal polysaccharides. As a result, shorter sugar oligomers and monomers become increasingly available and allow fast-growing opportunistic bacteria with a broader substrate spectrum to grow. Differences in nutritional strategies were apparent even between taxonomic classes; for example, in the expression of transport systems for nutrient uptake (Fig. 4A).

TonB-dependent transporter (TBDT) components dominated expressed transport proteins in *Flavobacteria*, whereas adenosine triphosphate (ATP)–binding cassette (ABC), tripartite ATP-independent periplasmic (TRAP), and tripartite tricarboxylate transporters (TTT) for low-molecular-weight (LMW) substrates were expressed only at low levels (Fig. 4A). TBDTs, originally thought to be restricted to complexed iron(III) (*17*) and vitamin B12 uptake, allow uptake of compounds that exceed the typical 600- to 800-dalton substrate range of normal porins (*18, 19*). Within *Bacteroidetes*, TBDTs are often colocalized with carbohydrate degradation modules (fig. S6) (*16*, *20*–*22*), and thus the substrate spectrum of these transporters may be much wider than anticipated (*23*), including oligosaccharides. TBDTs constituted no less than 13% of the expressed proteins identified during the bacterioplankton bloom at 31 March but only 7% in a non bloom sample at 11 February (fig. S7). This observation highlights the importance of TBDTs and corroborates a report of high TBDT expression in a coastal upwelling zone (*24*). In high-NPP zones, the capacity to take up oligomers as soon as they become transportable may constitute a major advantage over competitors restricted to smaller substrates.

In the *Gammaproteobacteria*, SAR92 featured a similar transporter expression profile as the *Flavobacteria*, whereas *Reinekea* spp. exhibited high expression of ABC and, to a lesser extent, TRAP transporters, indicating a different nutritional strategy with emphasis on the uptake of monomers (Fig. 4A).

Likewise, *Alphaproteobacteria* showed high expression levels of ABC and TRAP transporters and low levels of TBDTs and TTTs. This reflects the ecological strategy of the dominating SAR11. The well-studied representative *Pelagibacter ubique* HTCC 1062 thrives under oligotrophic conditions by means of high-affinity ABC and TRAP transporters and a constitutively expressed energy-producing proteorhodopsin (*25*–*27*). Our data confirmed constitutive proteorhodopsin expression and transporter components as the most abundant expressed proteins in the SAR11 clade, which corroborates previous findings (*28*). Members of the metabolically diverse, opportunistic alphaproteobacterial *Roseobacter* clade (*29*–*31*) exhibited LMW transporter expression levels that exceeded those of SAR11 (Fig. 4A). Although *Roseobacter* clade cells were two to four times less abundant than SAR11, they are larger, which may explain greater *Roseobacter* transporter expression.

Multiple factors may contribute to bacterioplankton bloom termination, such as predation by flagellate protozoa, viral lysis, and nutrient depletion. Phosphate limitation can spur algal exudate production, which might serve to promote the growth of phycosphere bacteria that remineralize and acquire phosphate more effectively (*32*); however, under phosphate limitation, algae and bacteria will compete. Phosphate dropped below the detection limit early in the phytoplankton bloom (fig. S1C), and the expression of several phosphate and phosphonate ABC-type uptake

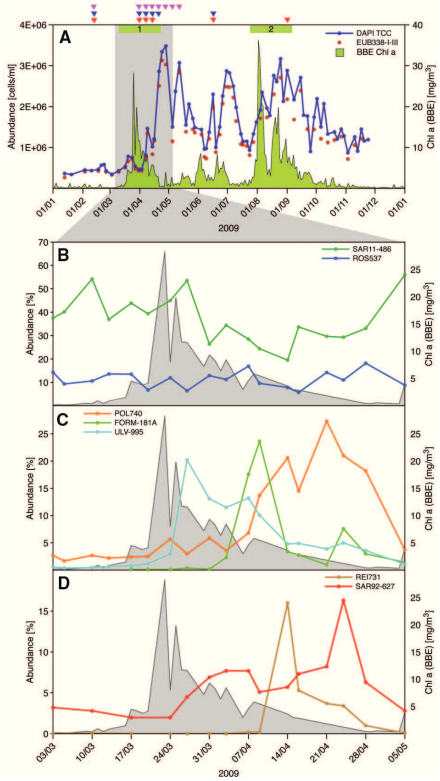


Fig. 1. Abundances of major bacterial populations during the bacterioplankton bloom as assessed by CARD-FISH. (**A**) Chlorophyll a (Chl a) concentration (measured with a BBE Moldaenke algal group analyzer), 4′,6-diamidino-2-phenylindole (DAPI)–based total cell counts (TCC), and bacterial counts (probe EUB338 I-III) during the year 2009; diatom-dominated spring blooms (1) and dinoflagellate-dominated summer blooms (2) are marked with green boxes; triangles on top mark accessory samples: metagenomics (red), metaproteomics (blue), and 16*S* rRNA gene tag sequencing (magenta). (**B**) Relative abundances of selected *Alphaproteobacteria*: SAR11 clade (probe SAR11-486) and *Roseobacter* clade (probe ROS537). (**C**) Relative abundances of selected *Flavobacteria*: *Ulvibacter* spp. (probe ULV-995), *Formosa* spp. (probe FORM-181A), and *Polaribacter* spp. (probe POL740). (**D**) Relative abundances of selected *Gammaproteobacteria*: *Reinekea* spp. (probe REI731) and SAR92 clade (probe SAR92-627). Further probes that are not shown for clarity are specified in the supplementary materials (tables S2 and S3).

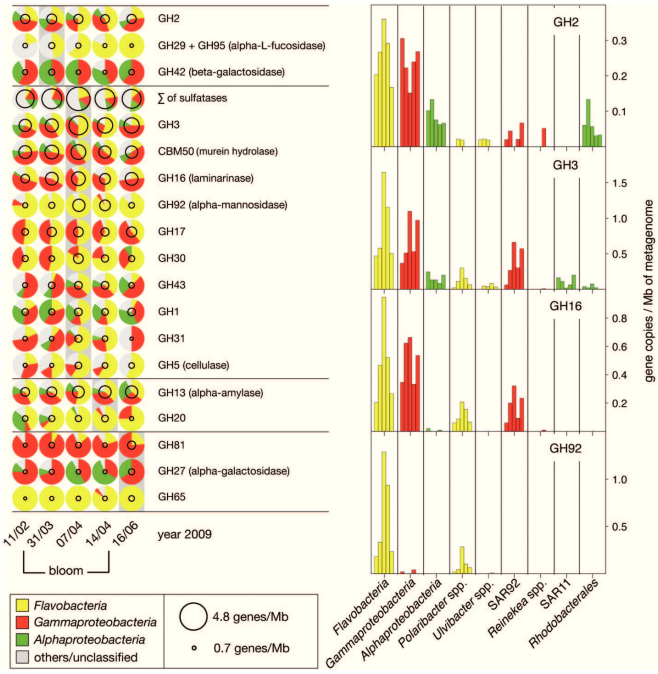


Fig. 2. Abundances of CAZymes with relevance for external carbohydrate degradation. (**Left**) Copies of 20 CAZymes per megabase of metagenome sequence with class-level taxonomic classifications (*12*). Maximum abundances are highlighted in gray. (**Right**) Detailed taxonomic breakdown for four selected CAZymes showing differing taxonomic compositions; each histogram shows data for the five metagenome samples (from left to right: 11 February 2009, 31 March 2009, 7 April 2009, 14 April 2009, and 16 June 2009).

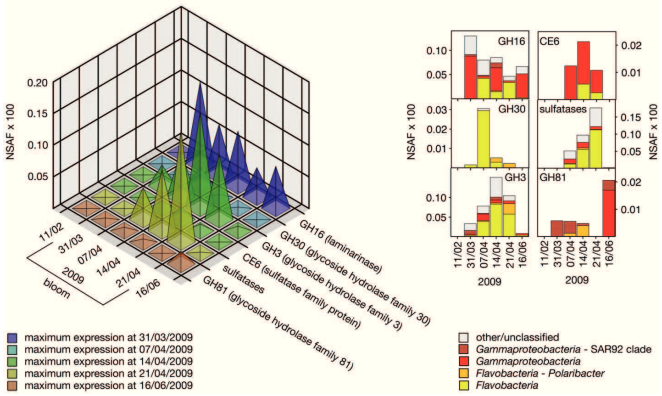


Fig. 3. Expression of CAZymes with relevance for external carbohydrate degradation; the proteome data were analyzed in a semiquantitative manner based on normalized spectral abundance factors (NSAFs) (*12*).

systems in various bacterial taxobins increased over the progression of the bloom (Fig. 4B). *Gammaproteobacteria* and SAR11 tended to use ABC-type phosphate transporters, as discovered in earlier studies (*28*), whereas flavobacterial *Polaribacter* spp. used phosphate:sodium symporters, and alphaproteobacterial *Rhodobacterales* spp. used phosphonate transporters.

In the first response to the phytoplankton bloom, flavobacterial *Ulvibacter* and *Formosa* spp. dominated (tables S2 and S4). Within these clades, TBDT components were among the proteins with the highest expression levels. This corroborates reports that specific *Flavobacteria* are tightly coupled to diatoms (*7*). *Bacteroidetes* have also been identified as major bacteria attached to marine snow (*33*, *34*), which agrees with their presumed role as fast-growing r strategists with specialization on the initial attack of highly complex organic matter (*16, 21, 35*). Hence, algal blooms lead to a multifold increase of colonization surfaces for *Bacteroidetes*, which respond with increased production of exoenzymes (*36*). After algal lysis, *Bacteroidetes* are the first to profit.

The second phase of the bacterioplankton succession coincided with a shift in algal composition (fig. S3) and was characterized by a pronounced peak of gammaproteobacterial *Reinekea* spp. that reached up to 16% of the bacteria (14 April 2009). *Reinekea* spp. featured a different expression profile, with high expression levels of transporters for peptides, phosphate, monosaccharides, and other monomers. These in situ data agree with the studies on cultured *Reinekea* species (*37*–*39*) that found broad generalist substrate spectra. The increase of alphaproteobacterial

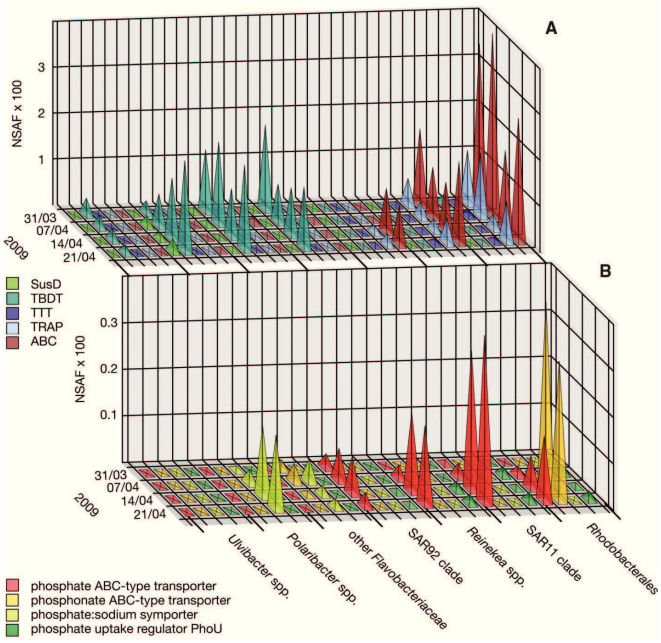


Fig. 4. Transporter components and phosphorus acquisition proteins of dominant taxa during the bacterioplankton bloom. (**A**) Expression of transporter components: starch utilization SusD-family proteins (SusD), TBDTs, TTTs, TRAPs, and ABCs. (**B**) Expression of proteins involved in phosphorus acquisition.

Roseobacter clade RCA during this phase might also be attributed to the *Roseobacter*'s opportunistic life-style (*29*) and is consistent with previous findings of free-living RCA phylotypes in the German Bight during diatom blooms (*40*).

The third phase of the spring 2009 bacterioplankton succession was dominated by flavobacterial *Polaribacter* and gammaproteobacterial SAR92 clade species, together with a secondary spike in *Formosa* spp. (Fig. 1, C and D). At this time, *Polaribacter* and *Formosa* dominated the particle/algae-attached fraction (table S8). Hence this phase with high sulfatase expression (Fig. 3) reflected another change of ecological niches (*12*).

Taken together, the bacterial response to coastal phytoplankton blooms was more dynamic than previously anticipated and consisted of a succession of distinct populations with distinct functional and transporter profiles. Thus, the diatom-induced growth of specific bacterioplankton clades most likely resulted from the successive availability of different algal primary products (bottom-up control), which provided the series of ecological niches in which specialized populations could bloom. As a result, we are now beginning to uncover the relevant predictors for defining the ecological niches of planktonic species (*41*) and thus can tackle the "paradox of the plankton" (*42*), which is how these species evade extinction by direct competition in a seemingly homogeneous habitat with limited resources.

References and Notes

1. J. P. Gattuso, M. Frankignoulle, R. Wollast, *Annu. Rev. Ecol. Syst.* **29**, 405 (1998).
2. C. B. Field, M. J. Behrenfeld, J. T. Randerson, P. Falkowski, *Science* **281**, 237 (1998).
3. F. Azam, *Science* **280**, 694 (1998).
4. F. Azam *et al.*, *Mar. Ecol. Prog. Ser.* **10**, 257 (1983).
5. J. Pinhassi *et al.*, *Aquat. Microb. Ecol.* **17**, 13 (1999).
6. L. Riemann, G. F. Steward, F. Azam, *Appl. Environ. Microbiol.* **66**, 578 (2000).
7. J. Pinhassi *et al.*, *Appl. Environ. Microbiol.* **70**, 6753 (2004).
8. J. M. Rinta-Kanto, S. Sun, S. Sharma, R. P. Kiene, M. A. Moran, *Environ. Microbiol.* **14**, 228 (2012).
9. L. B. Fandino, L. Riemann, G. F. Steward, R. A. Long, F. Azam, *Aquat. Microb. Ecol.* **23**, 119 (2001).
10. W. W. Lau, R. G. Keil, E. V. Armbrust, *Appl. Environ. Microbiol.* **73**, 2440 (2007).
11. Y. Tada *et al.*, *Appl. Environ. Microbiol.* **77**, 4055 (2011).
12. Further information is available as supplementary materials on *Science* Online.
13. B. L. Cantarel *et al.*, *Nucleic Acids Res.* **37** (Database issue), D233 (2009).
14. B. A. Wustman, M. R. Gretz, K. D. Hoagland, *Plant Physiol.* **113**, 1059 (1997).
15. V. B. Khodse, N. B. Bhosle, *Biofouling* **26**, 527 (2010).
16. P. R. Gómez-Pereira *et al.*, *Environ. Microbiol.* **14**, 52 (2012).
17. V. Braun, K. Hantke, *Curr. Opin. Chem. Biol.* **15**, 328 (2011).
18. T. K. Rostovtseva, E. M. Nestorovich, S. M. Bezrukov, *Biophys. J.* **82**, 160 (2002).
19. K. D. Krewulak, H. J. Vogel, *Biochem. Cell Biol.* **89**, 87 (2011).
20. M. Bauer *et al.*, *Environ. Microbiol.* **8**, 2201 (2006).
21. F. Thomas, J. H. Hehemann, E. Rebuffet, M. Czjzek, G. Michel, *Front. Microbiol.* **2**, 93 (2011).
22. B. M. Hopkinson, K. A. Barbeau, *Environ. Microbiol.* 10.1111/j.1462-2920.2011.02539.x (2011).
23. K. Schauer, D. A. Rodionov, H. de Reuse, *Trends Biochem. Sci.* **33**, 330 (2008).
24. R. M. Morris *et al.*, *ISME J.* **4**, 673 (2010).
25. S. J. Giovannoni *et al.*, *Science* **309**, 1242 (2005).
26. C. R. Reisch *et al.*, *Nature* **473**, 208 (2011).
27. J. Sun *et al.*, *PLoS ONE* **6**, e19870 (2011).
28. S. M. Sowell *et al.*, *ISME J.* **3**, 93 (2009).
29. M. A. Moran *et al.*, *Appl. Environ. Microbiol.* **73**, 4559 (2007).
30. T. Brinkhoff, H. A. Giebel, M. Simon, *Arch. Microbiol.* **189**, 531 (2008).
31. R. J. Newton *et al.*, *ISME J.* **4**, 784 (2010).
32. J. Tittel, O. Büttner, N. Kamjunke, *J. Plankton Res.* **34**, 102 (2012).
33. E. F. DeLong, D. G. Franks, A. L. Alldredge, *Limnol. Oceanogr.* **38**, 924 (1993).
34. D. Woebken, B. M. Fuchs, M. M. Kuypers, R. Amann, *Appl. Environ. Microbiol.* **73**, 4648 (2007).
35. J. L. Edwards *et al.*, *Genes* **1**, 371 (2010).
36. C. Arnosti, *Annu. Rev. Mar. Sci.* **3**, 401 (2011).
37. L. A. Romanenko, P. Schumann, M. Rohde, V. V. Mikhailov, E. Stackebrandt, *Int. J. Syst. Evol. Microbiol.* **54**, 669 (2004).
38. J. Pinhassi *et al.*, *Int. J. Syst. Evol. Microbiol.* **57**, 2370 (2007).
39. A. Choi, J. C. Cho, *Int. J. Syst. Evol. Microbiol.* **60**, 2813 (2010).
40. H. A. Giebel *et al.*, *ISME J.* **5**, 8 (2011).
41. S. J. Giovannoni, K. L. Vergin, *Science* **335**, 671 (2012).
42. G. E. Hutchinson, *Am. Nat.* **95**, 137 (1961).

Acknowledgments: We thank T. Hammer and T. Ferdelman for critical reading of the manuscript; M. Meiners, E. Karamehmedovic, B. Voigt, and V. Damare for sample processing; F. Ruhnau and L. Sayavedra for work on transporters; M. Zeder for automated counting; and R. Hahnke and J. Harder for help with probe testing. We are also grateful to our colleagues from the Bundesamt für Seeschifffahrt und Hydrographie for provision of operational model output. Analyses and visualizations used in fig. S1, D to F, were produced with the Giovanni online data system, developed and maintained by the NASA Goddard Earth Sciences Data and Information Service Center. We acknowledge the Moderate Resolution Imaging Spectroradiometer mission scientists and associated NASA personnel for these data. The sequence data reported in this study can be obtained from the European Bioinformatics Institute (study number ERP001227; www.ebi.ac.uk/ena/data/view/ERP001227). The German Federal Ministry of Education and Research (BMBF) supported this study by funding the Microbial Interactions in Marine Systems project (MIMAS, project 03F0480A, http://mimas-project.de).

Supplementary Materials

www.sciencemag.org/cgi/content/full/336/6081/608/DC1
Materials and Methods
Supplementary Text
Figs. S1 to S12
Tables S1 to S9
References (*43*–*102*)
Movie S1

22 December 2011; accepted 16 March 2012
10.1126/science.1218344

2.3 Carbohydrate-active enzyme profiling of a diatom-induced bacterioplankton succession reveals niche adaptations and trophic connections

Authors

Sixing Huang[1], Alexander J. Mann[1,2], Jost Waldmann[1,2], Richard Hahnke[1], Bernhard M. Fuchs[1], Jens Harder[1], Dörte Becher[3], Thomas Schweder[3], Frank Oliver Glöckner[1,2], Richard Reinhardt[4], Rudolf I. Amann[1], Hanno Teeling[1]*

[1] Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany
[2] Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen, Germany
[3] Institute of Marine Biotechnology e.V., Walther-Rathenau-Str. 49a, 17489 Greifswald, Germany
[4] Max Planck Genome Centre Cologne, Carl-von-Linné-Weg 10, 50829 Köln, Germany

*Corresponding author: Hanno Teeling

Publication status

In preparation

My contribution

Based on the results of the previous study [15], I further refined the CAZyme profiles of the four key players: [illegible] [illegible], *F*[illegible] spp., [illegible] spp. and [illegible] spp.. I compared the CAZyme gene frequencies of the MIMAS metagenomes with those from the GOS samples to highlight the abundant families. I identified the representative CAZyme families in each key player and investigated the corresponding substrates. Sulfatases and membrane transporters were also taken into consideration. I studied the physiological roles of these substrates in the hosting organisms and constructed a hypothetical model to explain the diatom-induced bacterioplankton succession. A molecular model for the degradation of algal polysaccharides in *F*[illegible] spp. was proposed. I also collected expression data and experimental results to verify the hypotheses.

Re□ea□ch a□ic□e □i□ □□e□a□a□i□□ f□□BMC Ge□□□ ic□

Ca□b□h□d□a□e-ac□i□e e□□□□e □□fi□i□g □f a dia□□□-i□d□ced bac□e□i□□a□□□□□ □□cce□□i□□ □e□ea□□ □iche ada□□a□i□□□ a□d □□□□hic c□□□ec□i□□□

Sixing Huang[1], Alexander J. Mann[1,2], Jost Waldmann[1,2], Richard Hahnke[1], Bernhard M. Fuchs[1], Jens Harder[1], D□rte Becher[3], Thomas Schweder[3], Frank Oliver Gl□ckner[1,2], Richard Reinhardt[4], Rudolf I. Amann[1], Hanno Teeling[1*]

[1] Max Planck Institute for Marine Microbiology, Celsiusstra□e 1, 28359 Bremen, Germany

[2] Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen, Germany

[3] Institute of Marine Biotechnology e.V., Walther-Rathenau-Str. 49a, 17489 Greifswald, Germany

[4] Max Planck Genome Centre Cologne, Carl-von-Linn□-Weg 10, 50829 K□ln, Germany

* Corresponding author: Hanno Teeling

Email addresses and telephone numbers of all authors:

Sixing Huang	shuang@mpi-bremen.de	+49 421 2028 928
Alexander J. Mann	amann@mpi-bremen.de	+49 421 2028 976
Jost Waldmann	jwaldman@mpi-bremen.de	+49 421 2028 907
Richard Hahnke	rhahnke@mpi-bremen.de	+49 421 2028 750
Bernhard M. Fuchs	bfuchs@mpi-bremen.de	+49 421 2028 935
Jens Harder	jharder@mpi-bremen.de	+49 421 2028 750
D□rte Becher	dbecher@uni-greifswald.de	+49 3834 864230
Thomas Schweder	schweder@uni-greifswalde.de	+49 3834 864212
Frank Oliver Gl□ckner	fog@mpi-bremen.de	+49 421 2028 970
Richard Reinhardt	reinhardt@mpipz.mpg.de	+49 221 5062-810
Rudolf I. Amann	ramann@mpi-bremen.de	+49 421 2028 930
Hanno Teeling	hteeling@mpi-bremen.de	+49 421 2028 976

Abstract

Background: Annually recurring spring phytoplankton blooms are typical for ocean margin zones in temperate latitudes. In a previous study, we investigated the response of marine subsurface bacterioplankton to a diatom-dominated spring phytoplankton bloom in the German Bight in 2009. Algal primary production triggered a bloom of bacterioplankton that was characterized by swift successions of dedicated bacterial clades with notably distinct substrate spectra. Prominent clades were the consecutively blooming bacterial genera *Formosa*, [illegible] and [illegible]. Here we present an analysis of these clades' carbohydrate-active enzymes with a focus on algal chrysolaminarin, callose, and sulfated polysaccharides as well as bacterial β-glucans.

Results: We re-analyzed metagenomic data from the spring bloom succession and amended these data with four newly sequenced draft-genomes of *Formosa* [illegible] (2), [illegible] (1) and [illegible] (1) isolates. In the early blooming *Formosa* [illegible] we found high gene frequencies of sulfatases, GH16 family glycoside hydrolases, β-glucanases, and glycosyltransferase family GT5 glycogen synthases. Metaproteome data confirmed [illegible] expression of GH16 and GT5 in *Formosa* [illegible], and *Formosa* [illegible] pure culture experiments demonstrated production of glycogen-like extracellular β-glucans. The intermediary and late blooming [illegible] and [illegible] had notably different carbohydrate-active enzyme profiles with high numbers of α-amylases, indicating α-glucan usage. [illegible] and [illegible] pure culture experiments verified α-glucan usage as well as chrysolaminarin usage in [illegible].

Conclusions: These findings suggest that *Formosa* [illegible] initiated the degradation of algal chrysolaminarin, callose, and sulfated polysaccharides and converted the excess to reserve and extracellular β-glucans. The initial algae degradation also exposed the intracellular chrysolaminarin to other microorganisms, paving the way for members of the clades [illegible] and [illegible]. These bacteria likely fed on more accessible less complex algal products, but also on the *Formosa* [illegible] β-glucans. Hence, carbohydrate-

active enzyme profiling could elucidate that the bacterioplankton succession was a bottom-up substrate-controlled displacement process of different bacterial clades with different niche adaptations that extended over multiple trophic levels and was underlain by intricate carbon conversions.

Ke□□ □□d□

bacterioplankton / CAZymes / diatoms / *F□□□ □□□* / metagenomics / microbial carbon pump / paradox of the plankton / phytoplankton bloom / □□□□□□□□□□□□□□/ □□□□□□□□□

Bac□g□□□□d

Land plants and marine algae each contribute about 50% of the global organic matter net primary production (NPP) [1]. In the oceans, diatoms are among the most important primary producers and are estimated to contribute about 20% to the global NPP [1, 2]. This culminates in annually recurring, sometimes massive diatom blooms that characterize the upwelling zones and continental shelves in higher latitudes worldwide [3-6].

Algal NPP provides not only the basis of the food web of marine micro-, meio- and macrofauna from protists to fish, but also of the heterotrophic bacterioplankton community. This trophic connection is reflected in synchronized blooms of bacteria after diatom blooms as they have been observed in various studies [6-13], and is largely mediated by algal polysaccharides. Polysaccharides can make up more than 50% of algal dry weight [14, 15], and many algae also produce polysaccharide-rich exudates, such as transparent exopolymer particles (TEPs). Diatom TEPs are adhesives rich in sulfated polysaccharides [16] that promote diatom flocculation and subsequent sinking [16, 17]. This removes diatoms from the photic zone and thereby accelerates bloom termination [16, 17]. Conversely, phycosphere bacteria feeding on TEP prolong diatom blooms in a mutualistic fashion [10]. Likewise, some free-living

marine bacteria respond to algal blooms by feeding on either living or decaying algal cells, which recycles diatom mineral components such as silica to the environment [18-21]. These primary diatom-degrading bacteria pave the way for more generalist bacteria and thus initiate a complex chain of nutrient utilizations among the heterotrophic bacterioplankton □ trophic relations as they are implied by the □microbial carbon pump□concept, which implies that microbes □pump□bioavailable organic carbon compounds into a pool of more recalcitrant compounds in a process that spans multiple trophic levels [22].

Polysaccharides constitute the most diverse class of macromolecules in nature, as they feature a wide variety of sugar monomers (including anhydro-sugars) that can be linked in various ways in a linear or branched fashion and decorated with a large number of moieties (e.g. sulfate, methyl or acetyl groups). In contrast proteins synthesized at ribosomes are usually linear and feature only peptide bonds, and the diversity of lipids is also limited. Hence the diversity and specificity of proteases and lipases is small in comparison to that of carbohydrate-active enzymes (CAZymes), i.e. enzymes that synthesize, modify or breakup glycosides [23, 24]. While there are bacterial species that can degrade most proteins and their amino acids or a large number of lipids, the diversity of carbohydrates is too large that a single bacterial species can harbor enough genes to break down even a larger proportion of the carbohydrate variants in nature. Thus, carbohydrate-degrading bacteria specialize on dedicated polysaccharides and their sugar monomers, which is reflected in the fact that CAZymes account for not more than about 2% of the genes in most bacterial genomes [23, 25]. This entails that many of such niche-adapted microbial specialist are required for the concerted break down of complex polysaccharides. A recent example is the study of Martens □□□□ [26], who showed this type of resource and niche partitioning in two human gut *B*□□□□□□□□□□□□□□.

In a previous in-situ study termed MIMAS (Microbial Interactions in Marine Systems) we reported on the bacterioplankton's response to a diatom-dominated phytoplankton bloom in the German Bight in spring 2009 [6]. This bloom started around the 2nd of March, peaked three weeks later around March 23rd and from there on subsided till the end of April. The diatoms were dominated by the genus Thalassiosira [6], centric diatoms of which the representative T. pseudonana has been extensively studied and completely sequenced [27]. A single Thalassiosira pseudonana cell is estimated to fix about 807.6 fg carbon per hour with sufficient sunlight under atmospheric CO_2 concentration (38 ppm) [28]. This corresponds to an average of 58 µg fixed carbon per liter and day when the Thalassiosira cell densities of the 2009 spring phytoplankton bloom in the German Bight [6] are integrated over the three main weeks. For a diatom bloom covering the 100 km x 100 km core area of the German Bight this would correspond to ~3 kilotons of carbon just in the topmost five meters of the water column. The phytoplankton bloom stirred a swift succession of distinct clades in the bacterioplankton (Figure 1). During the period of 20/03/2009 - 09/04/2009, relative abundances of *F*ormosa sp. (class *F*lavobacteria) rose quickly from below 1% to 23% of the bacterioplankton, with corresponding cell counts from 780 to 3.4 x 10^5 cells/ml. Afterwards, *F*ormosa relative abundances fell to below 10%. The genera Reinekea (class: *G*ammaproteobacteria) and Polaribacter (class *F*lavobacteria) responded later to the diatom decay. Reinekea relative abundances increased within one week from below 0.2% to 16% but vanished two weeks later with a relative abundance below 1%. At the same time, Polaribacter relative abundances rose from 10% to 20%. Several CAZymes, sulfatases, and sugar transporters were identified in the blooming bacterioplankton using a combination of metagenome and metaproteome analyses, but the substrates or metabolites remained elusive.

The aim of the current study was to gain more insight into the carbohydrate-mediated trophic relations among these key players of the bacterioplankton

succession. To this end we performed a series of comparative analyses for metagenomic data obtained during the bloom, and complemented these analyses with four newly sequenced draft genomes from the genera *F*□□□ □□□ (2), □ □□□□□□□□ (1) and □□□□□□□□□□□□ (1). In addition we performed cross-comparisons with the GOS metagenomic data [29], and thirteen published marine reference genomes. As a result, we hypothesize that members of *F*□□□ □□□ fed on complex algal β-1,3/β-1,6-linked glucans such as chrysolaminarin and callose. *F*□□□ □□□ converted these complex β-linked glucans to simpler □-linked glucans, which were among the substrates subsequently used by members of the clades □ □□□□□□□□ and □ □□□□□□□□□□□□. In addition, it could be shown that □□□□□□□□□□□□□were able to degrade chrysolaminarin. These findings were corroborated by metaproteome expression data and physiological data from the cultured isolates. In summary, this study exemplarily demonstrates that during bacterioplankton successions different bacterial clades with different niche adaptations can be functionally linked via intricate carbon compound conversion processes.

Re□□□□ a□d di□c□□□i□□

□□□ □□□□□□□□□□□□□□□□ □□ *07/04/2009* □ □□ □□□□□□□□□□□□□□ □□ □□□□ □□□□□□□□□ □□□ *GH16* □□□□ □□□□□□□□□□

Sulfatases are esterases that hydrolyze sulfate esters from carbohydrates, steroids and proteins. The disintegration of diatoms requires sulfatases since diatom TEP and cell wall casings both contain sulfated polysaccharides [16, 17]. A comparison of the sulfatase gene frequencies in metagenomes from the MIMAS and Global Ocean Survey (GOS) studies [29] revealed average frequencies of 0.13 □0.04% (Figure. 2). The MIMAS metagenome of 07/04/2009 had a sulfatase gene frequency of more than 3.5 standard deviations above the mean (0.28%), which constituted the highest sulfatase gene frequency in this comparison.

Diatoms synthesize two important □-1,3/□-1,6-glucans: chrysolaminarin and callose. Chrysolaminarin acts as primary storage compound [30], and callose is part of the gasket between the two diatom silica frustules [31] and might be involved in bio-silicification [32]. Both of these polysaccharides contain mostly □-1,3-linkages and can be degraded by □-1,3-glucanases of the glycoside hydrolase family 16 (GH16) [33-35]. On the other hand, □-1,6-linked glucans can be degraded by GH5 and GH30 □-1,6-glucanases [36]. GH16 family members are frequently found among marine and human gut *B*□□□□□□□□□□□□ and are involved in the degradation of various polysaccharides from marine algae such as sulfated polysaccharides like porphyran [37], □-carrageen and keratan sulfate. Comparative GH16 gene frequency analysis of the MIMAS and GOS metagenomes revealed that only three metagenomes exhibited levels above twice the mean of 0.023 □ 0.018% (Figure 2). All of these metagenomes belonged to the MIMAS study, with the highest level at 07/04/2009 (0.08%), followed by the preceding 31/03/2009 metagenome (0.06%) and the succeeding 14/04/2009 metagenome (0.06%). Thus, GH16 gene frequencies were elevated within the microbial community during this period of the diatom bloom. The 84 GH16 hits in the 07/04/2009 metagenome comprised 3 laminarinases, 46 □-glucanases, and 38 genes without substrate-specific annotations. Among the 46 □-glucanases, 30 were endo-□-1,3-glucanases. These □-glucanases genes were frequently co-localized with other CAZymes such as GH17 and the carbohydrate-binding domains CBM4, and occasionally CBM56. The GH17 family comprises members that are involved in the hydrolysis of □-1,3 linkages in laminarin and lichenin [38] and the turnover of callose [39], and the prokaryotic CBM4 and CBM56 are known to bind □-1,3-glucans and modulate the catalytic activity of functionally coupled glycoside hydrolase domains [40-43]. Finally, relatively low numbers of GH5 (0.03%) and GH30 (0.04%) were observed in 07/04/2009 metagenome probably due to the relatively low ratio of □-1,6 versus □-1,3 in chrysolaminarin (11:1) and in laminarin (15:1) [30].

F[illegible] GH16 [illegible] 07/04/2009

The aforementioned surge of sulfatases and GH16 in the MIMAS 07/04/2009 metagenome coincided with the bloom of *F[illegible]*, particularly *F[illegible]* (Figure 1). Both, taxonomic classification of the metagenomes into taxonomically coherent sequence bins (hereafter [illegible]taxobins[illegible]) [6, 44] and newly draft sequenced *F[illegible]* isolates supported that the *F[illegible]* clade caused these high sulfatases and GH16 frequencies. Draft genome sequencing of two *F[illegible]* isolates from the German Bight (*F[illegible]* sp. Hel3_A1_48 and *F[illegible]* sp. Hel1_33_131, hereafter referred to as *F[illegible]* group A and B) revealed that these two *F[illegible]* genomes were enriched in sulfatases and GH16 in comparison to other sequenced *B[illegible]* (Figure 3A). Corresponding gene frequencies were 0.4% to 0.9% for sulfatases and 0.3% for GH16, which consistently exceeds the levels of the 07/04/2009 metagenome. In contrast, other abundant taxa at this time such as the alphaproteobacterial SAR11 and [illegible] clades (see [6] for details) had only low sulfatase and GH16 gene frequencies. The genome of *C[illegible]* Pelagibacter ubique HTCC1062 of the SAR11 Ia clade contains neither sulfatases nor GH16 genes. Likewise the genomes of the [illegible] clade members *[illegible]. [illegible]* OCh 114 and *[illegible]. [illegible]* Och 149 possess only low sulfatase numbers and no GH16 genes. The gammaproteobacterial SAR92 and [illegible] clades, represented here by [illegible] HTCC2207, *[illegible]. [illegible]* MED297^T and the newly draft-sequenced *[illegible]* sp. Hel1_31_5_D35 (hereafter referred to as *[illegible]* sp. D35) were far less abundant than *F[illegible]* on 07/04/2009 (Figure 1) and thus could not account for the high sulfatase and GH16 gene frequencies, even though the respective genomes harbor moderate sulfatase and high GH16 gene numbers. This was corroborated by the 07/04/2009 metagenome, whose SAR92 taxobin yielded only nine contigs with

GH16 and thirteen contigs with sulfatase genes, and whose □ □□□□□□□□ taxobin even comprised no such contigs. Conversely, the *F□□□ □□□* taxobin from the MIMAS 07/04/2009 metagenome had an almost four times higher sulfatase and GH16 gene frequencies than the complete 07/04/2009 metagenome (Figure 4). The metagenome-derived *F□□□ □□□* gene frequencies were 0.3% for GH16 and 1% for sulfatases, which is within the ranges of the two *F□□□ □□□* draft-genomes. Likewise, absolute read count analysis of the metagenomes confirmed high abundances of *F□□□ □□□* sulfatases and GH16 family glycoside hydrolase genes. The *F□□□ □□□* taxobin contained 5,493 reads annotated as sulfatases and 1,348 reads as belonging to GH16, which exceeded the *G□□ □ □□□□□□□□□□□□□□□□* tenfold in sulfatases and sevenfold in GH16 genes (Suppl. Figure 1 A).

□□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□ □□□□□□□□□□□□□□□□□□□□ □□ *F□□□ □□□*

Marine *F□□□□□□□□□□□□* seem to have adapted to preferentially grow on macromolecules [45-48] and often contain high copy numbers CAZymes and sulfatases in conjunction with high numbers of TonB-dependent receptors, but the degradation process itself is only poorly understood. Sulfatases in bacteria have been found in the cytoplasm, periplasmic and extracellular space and remove the sulfate moieties from sulfated di- and monosaccharides resulting from the initial breakdown of sulfated polysaccharides [49-51]. GH16 gene products are not confined to the cytosol either [52]. Among the more than 1,300 reads identified as GH16 in the *F□□□ □□□* 07/04/2009 taxobin, 14% had secretion signal peptides. Although *F□□□ □□□* possessed high sulfatase and GH16 gene frequencies, this degradation process certainly involves more genes. As we reported before, *F□□□□□□□□□□□* expressed high levels of SusD and TonB transporter components during the bacterioplankton bloom [6]. SusD is part of a complex involved in nutrient binding on the cell surface [53-55]. The TonB system includes the TonB-dependent receptor Plug, TonB, ExbB and ExbD proteins [56, 57], and is responsible

for high-affinity import of molecules exceeding 600 Da across the outer membrane, including sugar oligomers [58]. This is achieved by a conformational change of the TonB pore that likely causes the plug domain to retract. The energy for this change stems from the proton motive force across the cytoplasmic membrane in a so far unknown mechanism [56] via the TonB protein that links the outer membrane complex to the cytoplasmic membrane. It was also reported that in *B*□□□□□□□□□□□□ □□□□□□□□□□□□ □□□□□□ ATCC 29148 and □□□□□□□□□□□□sp. MED152, SusD and TonB systems are working in concert to import molecules such as exogenous starch [26, 58-61]. Small substrates in the periplasmic space can be imported via ABC transporters at the expense of ATP [62]. Hence, initial diatom degradation by *F*□□□□□□□□□□□□ likely involves secreted degrading enzymes that initiate the extracellular degradation of algal polysaccharides to oligomers. These are then bound by the SusD domain on the cell surface and taken up via TonB-dependent transporters to the periplasm, where some desulfatation and a breakdown to di- and monomers takes place. The di- and monomers are subsequently transported via ABC transporters over the inner membrane into the cytoplasm, where remaining sulfate groups are removed (Suppl. Figure 2).

If the components of substrate-binding, outer membrane import, desulfation and glycoside hydrolysis are coordinately involved in the degradation of algal polysaccharides, positive correlation of their gene frequencies is expected. Indeed, frequencies of the Pfam profiles for TonB-dependent receptors (TonB_dep_Rec) and SusD in the *B*□□□□□□□□□□□□ reference genomes and the MIMAS metagenomes were positively correlated in an almost linear fashion (Figure 5A), probably due to the fact that they interact to form a complex on the outer membrane [59, 60]. Sulfatases and GH16 were positively related in an exponential manner (Figure 5B), although we have no evidence that sulfatases and GH16 are targeting at the same polysaccharide during the diatom degradation.

Three polysaccharide utilization loci (PULs) from the MIMAS 07/04/2009 metagenome corroborated these correlations. PULs are operon or regulon-like structures that are typically composed of glycoside hydrolases, carbohydrate-binding proteins, sulfatases and components of TonB-dependent transporters. All three PULs contained GH16 genes, TonB system components and SusD domain genes (Figure 6). Two of them likely stemed from *F*ormosa and the other one could be classified as of flavobacterial origin. These PULs bear similarities with the operons in the marine flavobacterium Zobellia galactanivorans Dsij^T [37] and the human gut isolate *B*acteroides plebeius DSM 17135 [37]. It is noteworthy that the PULs in Z. galactanivorans Dsij^T and *B*. plebeius DSM 17135 contain sulfatase genes, whose products are probably involved in removing the sulfate groups from porphyran. The absence of sulfatase in all the three putative PULs from the MIMAS 07/04/2009 metagenome indicated that sulfatases and GH16 do not directly collaborate on the same polysaccharide. The fact that these operons shared similarities with the marine bacterium Z. galactanivorans and the human gut bacterium *B*. plebeius [26, 45] suggests that PULs with similar organizations are widespread within the phylum *B*acteroidetes .

*A*lpha -glucans could link the successions of *F*ormosa , Reinekea and Polaribacter during the *2009* North Sea spring bloom

Analysis of the CAZyme profiles of *F*ormosa , Polaribacter and Reinekea draft genomes and metagenome taxobins indicated that □-glucans such as glycogen might have acted as substrates that linked their succession. Alpha-glucans do not contain the often-limiting elements nitrogen and phosphorus, so microbes can store excess carbon in glycogens when these elements are depleted, as during a bloom where algae and bacteria compete for nitrogen, phosphorus and iron. Hence, □-glucans serve primarily as storage compounds in prokaryotes during the stationary growth phase [1, 63].

However, it was also observed in □□□□□□□ □□□□ □□□□□□□□□□ CIP69.13 that extracellular glycogens are synthesized even when the cells are not subjected to nutrient limitations [64]. This finding also suggests other functions for □-glucans, probably forming a slime-like structure that coats the cell. Unlike plants and *C*□□□□□□□□□□□, diatoms use neither starch nor sucrose as main storage carbohydrates, but chrysolaminarin. Hence, if glycogens were produced, it is most likely from the blooming bacteria such as *F*□□□ □□□.

The glycoside transferase family GT5 comprises enzymes for □-glucan biosynthesis [1, 65], whereas families GH13, GH31, GH57 and GT35 families comprise enzymes for □-glucan degradation. GH13 and GH31 constitute diverse enzymes with various activities [66, 67]. Although GH13 includes enzymes such as □-1,3-glucan synthases and eukaryotic amino acid transporters, over 90% of the enzyme functions of its 35 GH13 subfamilies are related to □-glucan degradation (e.g. □-amylases, pullulanases, and □-glucosidases) [66]. GH31 members have been found in soil [67], human [68], plant [69] and marine bacteria [67] such as the flavobacterium *G*□□□ □□□□ □□□□□□□□KT0803 [70]. GH31 members catalyze the turnover of □-linked polysaccharides such as xyloglucans and most notably □-glucans. Members from GH57 have often been found in thermophile prokaryotes and are considered as the heat-stable versions of □-amylases [71-73]. Thus, GH57 is typically under-represented in temperate environments and consequently GH57 gene frequencies were below 0.01% in all nine MIMAS metagenomes. The only known activity of GT35 is that of glycogen phosphorylase (EC 2.4.1.1), which catalyzes releases glucose-1-phosphate from glycogen.

GT5 gene frequencies of *F*□□□ □□□ group A and B indicating □-glucan biosynthesis exceeded the metagenomic level on 07/04/2009 and were the highest among all the *B*□□□□□□□□□□□□ reference genomes (data not shown). As for subsequent glycogen degradation, □ □□□□□□□ □□□□□□□□□□ MED297^{T}, □ □□□□□□□ sp. D35, and □ □□□□□□□□□□□ sp.

Hel1_33_49 (hereafter they are referred to as Polaribacter sp. 49), Polaribacter irgensii 23-P and P. MED152 all had higher GH13 and GH31 gene frequency levels than the 14/04/2009 metagenome, while *F*ormosa group A and B had GH13 and GH31 levels that were lower than the metagenomes (Figure 3B). Conversely, the 07/04/2009 Reinekea taxobin did not contain any GT5 reads. In fact, 07/04/2009 Reinekea abundances were so low that only six corresponding CAZymes could be detected at all. *F*ormosa and Polaribacter on the other hand exhibited almost equally high GT5 frequencies. Of the 202 GT5 reads attributed to *F*ormosa , 66% were annotated as glycogen synthases (the others were annotated without substrate-specificities). On 14/04/2009, the Reinekea taxobin contained more than 2,500 GH13 reads (Suppl. Figure 2B) - about five times more than the Polaribacter taxobin - all with annotated glycogen degradation functions. However, the portion of GH13 with secretion signals was smaller in Reinekea (191 reads, 7%) than in Polaribacter (244 reads, 42%). Polaribacter also had more GH31 reads (Polaribacte : 349 reads from -glucosidases located on eight contigs; Reinekea : 183 reads from -glucosidases located one a single contig). None of the GH31 reads in both genera had secretion signals. *F*ormosa clade members on 14/04/2009 were poor in both, GH13 and GH31 on, probably due to their low abundance in that sample (Suppl. Figure 2B). There were 367 reads of GT35 in total on 14/04/2009 and the Reinekea taxobin accounted for half of these reads. In contrast, GT35 was absent in both, the Polaribacter taxobin and the Polaribacter reference genomes.

Taken together, Polaribacter and Reinekea had -glucan metabolic profiles complementary to *F*ormosa . The GT5 gene frequencies and read counts in *F*ormosa were high and those of GH13 and GH31 were low, whereas the opposite numerical trends were observed in Reinekea and Polaribacter in the genomes as well as the metagenome taxobins. These complementary profiles suggest that the studied *F*ormosa clade members produce more -glucans than they consume, whereas the

Reinekea and Polaribacter clade members consume more □-glucans than they produce.

It is interesting that though Reinekea and Polaribacter showed strong degradation capacities for □-glucans, there were subtle differences between their CAZyme profiles. First, Reinekea were equipped with more □-amylases and maltodextrin glucosidases. These enzymes first cleave large glycogen molecules into smaller oligosaccharides such as dextrin and maltose before cleavage into glucose monomers. Also, Reinekea encoded and expressed GT35 glycogen phosphorylases that were notably absent in Polaribacter Conversely, Polaribacter had more reads associated with □-glucosidases that release glucose units from the non-reducing ends one at a time [74, 75]. Without the fine granulation of glycogen by □-amylases, □-glucosidases can only catalyze the degradation in a slow manner. In comparison, □-amylases are much more efficient than □-glucosidases in degrading large glycogen molecules. Second, our previous study showed that on 14/04/2009 Reinekea expressed more ABC transporters, whereas Polaribacter expressed more TonB-dependent receptors but fewer ABC transporters [6]. The TonB system allows import of molecules larger than 600 Da and its lack in Reinekea limits □-glucan uptake to oligomers no longer than maltotriose (504 Da). As a result, Reinekea needed to deploy extracellular □-amylases to break down glycogen molecules to meet that constraint. In contrast, Polaribacter had less size constraints for outer membrane import. Furthermore, previous studies showed that the GH31 members were bound on the outer membrane [67, 76], suggesting that Polaribacter can degrade □-glucans extracellularly.

Polaribacter was also able to degrade chrysolaminarin

After the diatom cells were initially broken down by *F*ormosa at 14/04/2009, the once well-protected intracellular algal polysaccharides were exposed. This presented an opportunity for other heterotrophic bacteria to exploit the once inaccessible diatom

materials such as chrysolaminarin. As indicated previously in this report, CAZymes from GH16 are able to break down the β-1,3-linkages of the chrysolaminarin. The metagenome-derived □□□□□□□□□□□□taxobin at 14/04/2009 contained 568 reads of GH16. This value corresponded to 22% of all the identified GH16 genes in this metagenome and surpassed the value from then less abundant *F*□□□ □□□ (356 reads, 14%). Of all GH16 in the □□□□□□□□□□□□ taxobin, 223 (39%) were currently annotated as laminarinases. Two exemplary PULs were found in the □□□□□□□□□□□□ taxobin at 14/04/2009 (Figure 6) carrying genes for TonB-dependent receptors, SusD proteins and GH16 laminarinases. In contrast, the co-occurring genus □□□□□□□□ had only 61 reads of GH16 (2% of the total GH16 found in the 14/04/2009 metagenome).

While the □□□□□□□□ sp. D35 draft genome features only one GH16, the □□□□□□□□□□□□ sp. 49 draft genome has six GH16 genes (0.22%) (Figure 3A). These genes were annotated as laminarinases and β-glucanases. Two laminarinases are located within PULs in the draft genome (Figure 6). Together with the two PULs found in the □□□□□□□□□□□□taxobin at 14/04/2009, their organizations bear similarity with the ones found in other flavobacterial genomes such as *G*□□□ □□□□ □□□□□□□□KT0803 [70], *F*□□□□□□□□□□□□□□ □□□□□□□□□□□ UW101 [77] and *C*□□□□□□□□□□□□ □□□□□□ DSM 7489 [78]. As reported in the next section, the capability of utilizing laminarin was also confirmed experimentally in □□□□□□□□□□□□ sp. 49. In addition, it has already been confirmed that *F*□□□□□□□□□□□□□□ □□□□□□□□□□□ A3 was able to use laminarin [79], which is a strong indication that laminarin is a common substrate for various *F*□□□□□□□□□□□□□. Since chrysolaminarin shares the same types of glucosidic bonds as laminarin and both feature similar structures [15, 30], the ability to utilize laminarin could enable □□□□□□□□□□□□ sp. 49 to use also chrysolaminarin.

Being able to utilize laminarin, □□□□□□□□□□□□ could directly degrade the biomass from the diatoms and bypass intermediate consumers such as *F*□□□ □□□. In contrast, □□□□□□□□ likely could not have thrived without *F*□□□ □□□ to the observed densities.

Considering only about 10% of the energy is transmitted between two consecutive trophic levels, Polaribacter thus had a substantial competitive advantage against Reinekea by suffering less trophic losses along the stages of the food web. That could well be the reason why Polaribacter consistently outgrew Reinekea over the three weeks' time in the late bloom period (Figure 1). Although both *F*ormosa and Polaribacter could degrade diatom biomass, the Polaribacter bloom was delayed for a few days after the *F*ormosa bloom (Figure 1). This can be explained by the differences in their sulfatase gene levels. The 14/04/2009 Polaribacter taxobin had a sulfatase gene frequency of 0.2%, compared to the 1% of sulfatases from *F*ormosa s taxobin at 07/04/2009. Without sufficient sulfatases, the sulfated extracellular polymers could prevent Polaribacter from coming into contact with the diatom cells until they were removed by the *F*ormosa . In conclusion, Polaribacter could only bloom after *F*ormosa members, just as Reinekea did, even though they had different nutrient specializations.

Supporting information from the growth experiments and expression analyses

Re-mapping of the MIMAS study metaproteome expression data at 07/04/2009 onto the newly sequenced *F*ormosa group A and B draft genomes revealed that *F*ormosa group B indeed expressed GH16 glucanases as well as GT5 family glycogen synthases (Table 1). Conversely, sulfatases expression could be detected in the *F*ormosa taxobin at 07/04/2009 but not in the two draft genomes.

Microscopic observations showed that *F*ormosa group A cells were coated with EPS slime. When the cells were treated with -amylase, the EPS network began to disintegrate after 20 seconds. However, no EPS slime could be observed around *F*ormosa group B. This nevertheless fortifies the hypothesis that *F*ormosa did produce extracellular -glycans such as glycogens during the algal bloom.

Growth experiments were conducted to study the abilities of *F*ormosa group A and B, Reinekea sp. D35, Polaribacter sp. 49 to grow on minimal medium with glycogen or

laminarin as supplemented carbon source (Table 2). On minimal medium, the cells could grow up to 10^6 cells/mL. After supplementing laminarin, the cell densities of *F*□□□ □□□ group A and B continued to increase, but the supplemented glycogen did not have such an effect. On the other hand, the succeeding □ □□□□□□□ sp. D35 and □ □□□□□□□□□□□□□sp. 49 could grow on glycogen. In addition, □ □□□□□□□□□□□□□sp. 49 could also grow on laminarin. Together, these observations confirmed the hypotheses derived from □omics data.

C□□c□□□i□□□

The booming science of glycobiology is continuously expanding our knowledge on the mechanisms, substrate-specificities and phylogeny of various classes of CAZymes. This knowledge is reflected in the continuously growing Carbohydrate-Active EnZymes online database [24], which has propelled the successful application of CAZyme analysis in the field of environmental microbiology [26, 37, 80, 81].

This study extends the previous reported MIMAS study [6] and tries to uncover the mechanics that drove the successive blooms of diatoms, *F*□□□ □□□, □ □□□□□□□ and □ □□□□□□□□□□□□ in North Sea in spring 2009. In *F*□□□ □□□, we observed high gene frequencies of sulfatases, GH16 □-glucanases, and GT5 glycogen synthases, whereas we observed high frequencies of □-amylases within □ □□□□□□□ and □ □□□□□□□□□□□□ and high GH16 frequencies in □ □□□□□□□□□□□□. Diatom cells secret sulfated polysaccharides to build extracellular structures such as transparent exopolymer particles (TEP) [16, 17] and use □-glucans, especially □-1,3-glucans as storage and structural compounds [30, 31, 82]. Therefore we hypothesize that the *F*□□□ □□□ bacteria in our study were able to feed on the complex algal polysaccharides and synthesize simpler, linear polymers such as the □-linked glycogen. It was shown that heteropolymers with complex tertiary structures are less accessible for the general bacteria than the simple, linear homopolymers [83]. Thus, the □-glucan that *F*□□□ □□□ synthesized during its bloom

likely served as one of the substrates for the subsequent blooms of Reinekea and Polaribacter Chrysolaminarin from the disrupted diatom cells could also contribute to the bloom of the laminarin-degrading Polaribacter Meta-expression and physiological laboratory data support these findings. Although the microbial interactions investigated in this study must be considered as a small part of a larger and more complex carbon transformation event chain, the characterization of the substrates and their metabolic enzymes in the blooming taxa constitute first step into understanding such a complex natural process. While the details and true complexities of these relationships need further in-depth studies, we demonstrate that our methods can provide vital insights into aspects of the intricate interactions between diatoms and the blooming bacterioplankton and reveal a part of the underlying carbon transformation processes.

Methods

Formosa group A and B, Reinekea sp. D35, Polaribacter sp. 49

Formosa group A and B, Reinekea sp. D35, Polaribacter sp. 49 were isolated from a surface seawater sample from the German Bight (54° 11’ 03’’ N, 7° 54’ 00‘‘ E) by Richard Hähnke (unpublished). Except Formosa group A, which was collected on 03/09/2010, all the other isolates were sampled on 20/04/2010. Both Formosa group A and B are closely related with F. agariphila KMM 3901^T (97% 16S rRNA sequence similarity) [84]. Their similarities with F. algae KMM 3553^T [85] are 95%. Reinekea sp. D35 shows a high level of similarity with R. marinisedimentorum KMM 3655^T [86] and R. blandensis MED297^T (96 - 97% 16S rRNA sequence similarity) [87]. The closest relative of Polaribacter sp. 49 is P. irgensii 23-P (95% similarity) [88].

Bacterial genomes and metagenomes

The two North Sea isolates Formosa group A and B were draft sequenced using 454 FLX Ti technology (454 Life Sciences, Branford, CT, USA) by LGC Genomics (LGC

Genomics GmbH, Berlin, Germany). Within the genus *Polaribacter* , the genomes of *Polaribacter irgensii* 23-P [88] and *Polaribacter dokdonensis* MED152 [61] are available in NCBI BioProject. The German Bight isolates *Polaribacter* sp. 49 was also draft sequenced using on the 454 FLX Ti platform (454 Life Sciences) by LGC Genomics (LGC Genomics GmbH). Within the genus *Reinekea* , the genome of *Reinekea blandensis* MED297^T is published [87]. *Reinekea* sp. D35 from the North Sea was sequenced using PacBio RS technology by GATC Biotech (GATC Biotech AG, Konstanz, Germany). The draft genomes of our North Sea isolates are available from the ENA (accession numbers), and their basic properties are summarized in Supplementary Table 1.

Reference genomes were selected based on the abundant bacterial taxa identified in our previous study [6]. The set included *Candidatus Pelagibacter ubique* HTCC1062 [89, 90], *Roseobacter denitrificans* OCh 114 [91], *Roseobacter litoralis* Och 149 [92], the SAR92 clade *gamma proteobacterium* HTCC2207 [93] and *Ulvibacter* sp. SCB49 [94]. In addition, we also included a group of *Bacteroidetes* with known macromolecule degradation activities for comparison, including *Cellulophaga algicola* DSM 14237 [95, 96], *Cellulophaga lytica* DSM 7489 [78], *Flavobacterium johnsoniae* UW101 [77], *Zobellia galactanivorans* Dsij^T [37] and *Cytophaga hutchinsonii* ATCC 49512 [54].

The raw data of the eight metagenomes of our previous study are available from the European Bioinformatics Institute ENA archive (study number ERP001227). In addition to these metagenomes, a ninth metagenome from 01/09/2009 was also included in the current study. This metagenome comprised two PTPs of 454 FLX Ti sequencing (923,022 reads, assembly: 82,466 contigs, 79 Mb) and is also available from the ENA archive (insert number).

The GOS data set is a comprehensive survey over the global ocean surface water [29]. Seventeen GOS samples were compared in this study to provide an overview about gene frequencies in the global oceans. All these samples stem from 0.1 - 0.8 µm

filter fractions dominated by planktonic bacteria. Respective metagenomes were sampled from the Atlantic, the Pacific and the Indian Ocean and cover various habitats such as coral reefs, mangrove forests, estuaries, coastal regions and open oceans. Additional Global Ocean Survey (GOS) metagenomes were downloaded from the MG-RAST server [97] in FASTA format after the gene calling: GS001c, GS006, GS009, GS011, GS012, GS020, GS021, GS026, GS032, GS033, GS048a, GS048b, GS049, GS110b, GS112b, GS115, GS120.

D□□□ □□□□□□□□□□

The steps of read dereplication, assembly, annotation and partitioning of metagenomes into taxonomically coherent bins (□taxobins□) were carried out as previously described [6, 44] with two additional steps: First, the dereplicated sequence reads were remapped onto the assembled contigs, in order to calculate the read coverage of every gene. Second, the hidden Markov Models (HMM) of every viableCAZyme family was downloaded from the dbCAN server [98] and used to search the metagenomes. The results were filtered with an e-value cutoff of E-15 and used to complement our previous BLAST- and Pfam-based CAZy annotations. The draft genomes of *F*□□□ □□□ group A and B, □□□□□□□□ sp. D35, □□□□□□□□□□□□□sp. 49 were processed in a similar way without the read remapping step.

All GOS metagenomes and public single genomes were processed as follows: Protein sequences were searched against the CAZy database [24] as of 20/05/2011. For every protein, the best hit with an e-value below E-15 was kept. In addition, sequences were search against the Pfam v24 database. The reference single bacterial genomes were downloaded as protein FASTA files and processed as the GOS data, except when their CAZy annotations were already available on the CAZy/Genomes page (http://www.cazy.org/Genomes.html); then the official annotations were used.

[illegible]

In this study, we performed comparisons among multiple metagenomes as well as comparisons between metagenome taxobins and individual genomes. Comparisons among non-normalized absolute values are valid only within the same sample. In contrast, data must be normalized before cross-sample comparisons to account for the effect of different sample sizes. In the current study, both absolute and normalized values were used. Absolute values reflect the combined contribution of all individuals within a taxobin to the community and they are dependent on cell counts. We used these absolute quantities to describe the functional composition of the microbial community. In contrast, normalized values, such as gene copy numbers were used to elucidate the genotypes of the metagenomes and genomes. In this study, we used the following definition of gene frequency for the cross-sample comparison:

In contrast to absolute gene counts, gene frequencies reveal the characteristics of an individual taxobin and can be used for within as well as across sample comparisons. Moreover, gene frequencies do not depend on the cell abundances.

To compute the correlations between SusD domains and TonB-dependent receptors and between sulfatases and GH16 in the *B*[illegible] genomes, the flavobacterial taxobins from all nine metagenomes were extracted and the corresponding gene frequencies were calculated. Together with nine *B*[illegible] reference genomes (complete list is in the [illegible]*B*[illegible] section), the values were plotted and the fit functions were sought.

C[illegible] *F*[illegible] group A [illegible] *B*

Cells from the pure cultures of *F*[illegible] group A and B were treated with 33 U/mg of [illegible]-amylases. To visualize the treatment effects on EPS, *F*[illegible] cells were stained with

DAPI and FITC-labeled lectin WGA (Bennke In preparation) and visually examined under a fluorescence microscope.

E group A *B,* sp. D35 sp. 49

F group A and B, sp. D35 and sp. 49 were first cultivated in 200 M minimal medium Hel_minV. Hel_minV is a modified version described by Widdel and Bak [99]. The medium contains glucose, cellobiose, yeast extract, amino acids, vitamin, trace element and EDTA. The cells were only able to grow up to 10^6 cells/mL in this medium. *F* group A was kept at 25 C, while the other cultures were kept at 12 C.

Four to five flakes of laminarin from (Sigma-Aldrich L1760) were first washed with Milli-Q water. Afterwards, they were heated to 70 C for one hour and then washed again with Milli-Q. The process was repeated three times. 10 mL of the cleaned laminarin were added to each culture. After one to three weeks of incubation, the cultures were inspected for pellet formation in comparison with a laminarin-free control culture.

A glycogen stock solution (100 g/L Sigma-Aldrich G-8751) was first filtered through a 0.2 m sterile filter. Afterwards, 200 L stock solution was added to 10 mL of the cell cultures. After one to three weeks of incubation, the cell cultures were checked for pellet formation in comparison with a glycogen-free control culture.

List of abbreviations used

CAZyme,	carbohydrate-active enzymes
GH	glycoside hydrolase
GT	glycosyltransferase
EPS	extracellular polymeric substances

PUL	polysaccharide utilization locus (plural loci)
F□□□ □□□ group A	*F□□□ □□□* sp. Hel3_A1_48
F□□□ □□□ group B	*F□□□ □□□* sp. Hel1_33_131
□□□□□□□□ sp. D35	□□□□□□□□ sp. Hel1_31_5_D35
□□□□□□□□□□□□□□□ sp. 49	□□□□□□□□□□□□□□□ sp. Hel1_33_49
□□□□□□□□□□□□□□□ sp. 85	□□□□□□□□□□□□□□□ sp. Hel1_85
TEP	transparent exopolymer particles
GOS	Global Ocean Survey
EDTA	Ethylenediaminetetraacetic acid

A□□h□□□□c□□□ib□i□□□

SH carried out the CAZy analyses of the metagenomes and reference genomes interpreted the CAZy profiles and drafted the manuscript. AM and JW participated in the taxonomic classification of the metagenomes. RH and JH were responsible for the pure cultures and physiological experiments. DB and TS provided the metaproteomic data. BF, FOG and RA participated in the design and organization of the study. HT oversaw the project and wrote substantial parts of the manuscript. All the authors read and approved the final manuscript.

Ac□□□□ edg□ e□□

The German Federal Ministry for Research and Education (MIMAS project, No. 03F0480) and the Max Planck Society supported this study; we thank Clara Mart□nez P□rez for her genomic analysis of the three *F□□□ □□□* isolates; Johannes Werner for critical reading of the manuscript; we are also grateful to Christin Bennke for her work in visualizing the EPS of *F□□□ □□□*.

Refe□e□ce□

1. Field CB, Behrenfeld MJ, Randerson JT, Falkowski P: **Primary production of the biosphere: integrating terrestrial and oceanic components.** *Science* 1998, **281**:237-240.
2. Falkowski PG, Barber RT, Smetacek V: **Biogeochemical Controls and Feedbacks on Ocean Primary Production.** *Science* 1998, **281**:200-207.
3. Lignell R, Heiskanen AS, Kuosa H, Gundersen K, Kuuppo-Leinikki P, Pajuniemi R, Uitto A: **Fate of a phytoplankton spring bloom: sedimentation and carbon flow in the planktonic food web in the northern Baltic.** *Marine Ecology-Progress Series* 1993, **94**:239.
4. Shinada A, Shiga N, Ban S: **Origin of [illegible] diatoms that cause the spring phytoplankton bloom in Funka Bay, southwestern Hokkaido, Japan.** *Plankton Biology and Ecology* 1999, **46**:89-93.
5. Maier G, Glegg GA, Tappin AD, Worsfold PJ: **A high resolution temporal record of phytoplankton bloom dynamics in the eutrophic Tamar Estuary (SW England).** *Science of the Total Environment* 2012, **434**:228 - 239.
6. Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM, Kassabgy M, Huang S, Mann AJ, Waldmann J, Weber M, Klindworth A, Otto A, Lange J, Bernhardt J, Reinsch C, Hecker M, Peplies J, Bockelmann FD, Callies U, Gerdts G, Wichels A, Wiltshire KH, Glöckner FO, Schweder T, Amann R: **Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom.** *Science* 2012, **336**:608-611.
7. Cole JJ: **Interactions between bacteria and algae in aquatic ecosystems.** *Annual Review of Ecology and Systematics* 1982, **13**:291-314.
8. Bird DF, Kalff J: **Empirical relationships between bacterial abundance and chlorophyll concentration in fresh and marine waters.** *Canadian Journal of Fisheries and Aquatic Sciences* 1984, **41**:1015-1023.
9. Bell RT, Kuparinen J: **Assessing phytoplankton and bacterioplankton production during early spring in Lake Erken, Sweden.** *Appl Environ Microbiol* 1984, **48**:1221-1230.
10. Smith DC, Steward GF, Long RA, Azam F: **Bacterial mediation of carbon fluxes during a diatom bloom in a mesocosm.** *Deep Sea Research Part II: Topical Studies in Oceanography* 1995, **42**:75-97.
11. Riemann L, Steward GF, Azam F: **Dynamics of bacterial community composition and activity during a mesocosm diatom bloom.** *Appl Environ Microbiol* 2000, **66**:578-587.

12. Pinhassi J, Sala MM, Havskum H, Peters F, Guadayol O, Malits A, Marrasé C: **Changes in bacterioplankton composition under different phytoplankton regimens.** *Appl Environ Microbiol* 2004, **70**:6753-6766.

13. Niu Y, Shen H, Chen J, Xie P, Yang X, Tao M, Ma Z, Qi M: **Phylogenetic composition and succession of bacterioplankton communities in Lake Taihu, China.** *Microb Ecol* 2011, **45**:4169-4182.

14. Gatenby CM, Orcutt DM, Kreeger DA, Parker BC, Jones VA, Neves RJ: **Biochemical composition of three algal species fed as monocultures and mixed diets to juvenile unionid mussels.** *J Exp Mar Biol Ecol* 2003, **15**:1-11.

15. Alderkamp A-C, Van Rijssel M, Bolhuis H: **Characterization and isolation of marine bacteria and the activities of their exopolysaccharides during degradation of the algal storage glucan laminarin.** *FEMS Microbiol Ecol* 2006, **59**:108-117.

16. Alldredge AL: **The quantitative significance of gelatinous zooplankton as prey, predators, and competitors.** *Ophelia* 1999, **35**:397-400.

17. Passow U, Alldredge AL, Logan BE: **The role of particulate carbohydrate exudates in the flocculation of diatom blooms.** *Deep Sea Research I* 1994, **41**:335-357.

18. Hellebust JA: **Excretion of some organic compounds by marine phytoplankton.** *Limnol Oceanogr* 1965, **10**:192-206.

19. Mague TH, Friberg E, Hughes DJ, Morris I: **Extracellular release of carbon by marine phytoplankton; a physiological approach.** *Limnol Oceanogr* 1980, **25**:262-279.

20. Baines SB, Pace ML: **The production of dissolved organic matter by phytoplankton and its importance to bacteria: Patterns across marine and freshwater systems.** *Limnol Oceanogr* 1991, **36**:1078-1090.

21. Sarmento H, Gasol JM: **Use of phytoplankton-derived dissolved organic carbon by different types of bacterioplankton.** *Environ Microbiol* 2012, **14**:2348-2360.

22. Jiao N, Herndl GJ, Hansell DA, Benner R, Kattner G, Wilhelm SW, Kirchman DL, Weinbauer MG, Luo T, Chen F, Azam F: **Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean.** *Nat Rev Microbiol* 2010, **8**:593-599.

23. Henrissat B, Coutinho PM, Davies GJ: **A census of carbohydrate-active enzymes in the genome of Arabidopsis thaliana.** *Plant Mol Biol* 2001, **47**:55-72.

24. Cantarel BL, Coutinho PM, Rancurel C, Bernard T, Lombard V, Henrissat B: **The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics.** *Nucleic Acids Res* 2009, **37**:D233-D238.

25. Coutinho PM, Deleury E, Davies GJ, Henrissat B: **An evolving hierarchical family classification for glycosyltransferases.** *J Mol Biol* 2003, **328**:307-317.

26. Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, Abbott DW, Henrissat B, Gilbert HJ, Bolam DN, Gordon JI: **Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts.** *PLoS Biol* 2011, **9**:e1001221.

27. Armbrust EV, Berges JA, Bowler C, Green BR, Martinez D, Putnam NH, Zhou S, Allen AE, Apt KE, Bechner M, Brzezinski MA, Chaal BK, Chiovitti A, Davis AK, Demarest MS, Detter JC, Glavina T, Goodstein D, Hadi MZ, Hellsten U, Hildebrand M, Jenkins BD, Jurka J, Kapitonov VV, Kröger N, Lau WWY, Lane TW, Larimer FW, Lippmeier JC, Lucas S, Medina M, Montsant A, Obornik M, Parker MS, Palenik B, Pazour GJ, Richardson PM, Rynearson TA, Saito MA, Schwartz DC, Thamatrakoln K, Valentin K, Vardi A, Wilkerson FP, Rokhsar DS: **The genome of the diatom *Thalassiosira pseudonana*: ecology, evolution, and metabolism.** *Science* 2004, **306**:79-86.

28. Sobrino C, Ward ML, Neale PJ: **Acclimation to Elevated Carbon Dioxide and Ultraviolet Radiation in the Diatom "*Thalassiosira pseudonana*": Effects on Growth, Photosynthesis, and Spectral Sensitivity of Photoinhibition.** *Limnol Oceanogr* 2008, **53**:494-505.

29. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y-HH-HH, Smith HO: **Environmental genome shotgun sequencing of the Sargasso Sea.** *Science* 2004, **304**:66-74.

30. Beattie A, Hirst EL, Percival E: **Studies on the metabolism of the Chrysophyceae. Comparative structural investigations on leucosin (chrysolaminarin) separated from diatoms and laminarin from the brown algae.** *Biochem J* 1961, **79**:531-537.

31. Granum E, Myklestad SM: **A simple combined method for determination of β-1,3-glucan and cell wall polysaccharides in diatoms.** *Hydrobiologia* 2002, **477**:155-161.

32. Law C, Exley C: **Ne□ i□□igh□i□□□ □ilica de□□□i□i□□ i□ h□□□e□ai□ (□□□□□□□□□ □□□□□□□).** *B□ C □□□□□B□□□□□* 2011, **11**:112.

33. Strohmeier M, Hrmova M, Fischer M, Harvey AJ, Fincher GB, Pleiss J: **M□□ec□□a□ □□de□i□g □f fa□i□□ GH16 g□□c□□ide h□d□□□a□e□: □□□e□□ia□ □□□e□ f□□ □□□□g□ca□ □a□□g□□c□□□□□a□e□/h□d□□□a□e□ i□ ce□□ □a□□ □□difica□i□□ i□ □he □□aceae.** *□□□□□□ □□□* 2004, **13**:3200-3213.

34. Hrmova M, Fincher GB: **P□a□□ a□d □ic□□bia□ e□□□□□e□ i□□□□□ed i□ □he de□□□□□e□i□a□i□□ □f (1,3)-□-d-g□□ca□□ a□d □e□a□ed □□□□□acch a□ide□.** In *C□□□□□□, B□□□□□□□□ □□□ B□□□□□□ □□ (1,3)-□-□-G□□□□□□ □□□ □□□□□□ □□□□□□□□□□□□□□*. 1st edition. Academic Press, Elsevier Inc, San Diego, USA; 2009:119-170.

35. Nakajima M, Yamashita T, Takahashi M, Nakano Y, Takeda T: **A □□□e□ g□□c□□□□□h□□□ha□id□i□□□□i□□-a□ch□□ed g□□c□□ide h□d□□□a□e f□□□ U□□i□ag□ e□c□□e□□a f□□c□i□□□□ i□ □-1,3-g□□ca□ deg□ada□i□□.** *A□□□E□□□□□ □ □□□□□□* 2012, **78**:5682-5689.

36. St John FJ, Gonz□lez JM, Pozharski E: **C□□□□□□ida□i□□ □f g□□c□□□□□h□d□□□a□e fa□i□□ 30: a d□a□d□□ai□ 4/7 h□d□□□a□e fa□i□□ c□□□i□□i□g □f □□□ □□□□c□□□a□□ di□□i□c□g□□□□□.** *FEB□ □□□□* 2010, **584**:4435-4441.

37. Hehemann J-H, Correc G, Barbeyron T, Helbert W, Czjzek M, Michel G: **T□a□□fe□ □f ca□b□h□d□a□e-ac□i□e e□□□□□e□ f□□□ □a□i□e bac□e□ia □□ Ja□a□e□e g□□ □ic□□bi□□a.** *□□□□□* 2010, **464**:908-912.

38. Hreggvidsson GO, Dobruchowska JM, Fridjonsson OH, Jonsson JO, Gerwig GJ, Aevarsson A, Kristjansson JK, Curti D, Redgwell RR, Hansen C-EE-EE, Kamerling JP, Debeche-Boukhit T: **E□□□□□i□g □□□□e□ □□□-Le□□i□ □-g□□c□□□□□□a□□fe□a□e□ f□□□ □□□□e□bac□e□ia f□□ □□dif□i□g □i□ea□ (□1->3)-□i□□ed g□□c□-□□ig□□accha□ide chai□□.** *G□□□□□□□□□□* 2011, **21**:304-328.

39. Rinne PLH, Welling A, Vahala J, Ripel L, Ruonala R, Kangasj□rvi J, van der Schoot C: **Chi□□i□g □f d□□□a□□ b□d□ h□□e□i□d□ce□ □□□□ □□□□□ □□□□□ □ a□d □ec□□i□□ GA-i□d□cib□e 1,3-be□a-g□□ca□a□e□ □□ □e□□□e□ □ig□a□ c□□d□i□□ a□d □e□ea□e d□□□a□c□ i□ □□□□□□□.** *□□□□□C□□* 2011, **23**:130-146.

40. Boraston AB, Bolam DN, Gilbert HJ, Davies GJ: **Ca□b□h□d□a□e-bi□di□g □□d□□e□: fi□e-□□□i□g □□□□□accha□ide □ec□g□i□i□□.** *B□□□□□□ J* 2004, **382**:769-781.

41. Nordberg Karlsson E, Bartonek-Roxbackslashra E, Holst O: **C□□□i□g a□d □e□□e□ce □f a □he□□□□□ab□e □□□id□□ai□ □□□a□a□e f□□□ □he bac□e□i□□ □□□□□□□□□□□ □□□ □□□□□□.** *B□□□□□□ B□□□□□□A□□□* 1997, **1353**:118-124.

42. Zverlov VV, Volkov IY, Velikodvorskaya GA, Schwarz WH: **The bi□di□g □a□e□□ □f □□□ ca□b□h□d□a□e-bi□di□g □ □d□e□ □f □a□ i□a□i□a□e La□ 16A f□□□ □□□□□ □□□□□ □□□□□□□□□□: diffe□e□ce□ i□ be□a-g□□ca□ bi□di□g □i□hi□ fa□i□□ CBM4.** □ □□□□□□□□□□□□ 2001, **147**:621-629.

43. Yamamoto M, Ezure T, Watanabe T, Tanaka H, Aono R: **C-□e□□ i□a□d□□ ai□ □f be□a-1,3-g□□ca□a□e H i□ □□□□□□□ □□□□□□□□□ IAM1165 ha□ a □□e i□ bi□di□g □□ i□□□□□be□ be□a-1,3-g□□ca□.** *FEB□ □□□□* 1998, **433**:41-43.

44. Teeling H, Gl□ckner FO: **C□□e□□ □□□□□□□i□ie□ a□d cha□e□ge□ i□ □ic□□bia□ □e□age□□□e a□a□□□i□--a bi□i□f□□□ a□ic □e□□□ec□i□e.** *B□□□ B□□□□□□□□* 2012, **13**:728-742.

45. Thomas F, Hehemann J-H, Rebuffet E, Czjzek M, Michel G: **E□□i□□□□ e□□a□a□d g□□bac□e□□ide□e□: □he f□□d c□□□□ec□i□□.** *F□□□□□□ □□□□□□□□* 2011, **2**:93.

46. Cantarel BL, Lombard V, Henrissat B: **C□□ □□e□ ca□b□h□d□a□e □□i□□a□i□□ b□ □he hea□□h□h□□ a□ □ic□□bi□□ e.** □□□□ □ □□□ 2012, **7**:e28742.

47. Martinez-Garcia M, Brazel DM, Swan BK, Arnosti C, Chain PSG, Reitenga KG, Xie G, Poulton NJ, Lluesma Gomez M, Masland DED, Thompson B, Bellows WK, Ziervogel K, Lo C-CC, Ahmed S, Gleasner CD, Detter CJ, Stepanauskas R: **Ca□□□□i□g □i□g□e ce□□ ge□□□□e□ □f ac□i□e □□□□□□accha□ide deg□ade□□: a□ □□e□□ec□ed c□□□□ib□□i□□ □f □□□□□□□□□□ □□□□□□□□.** □□□□□ □ □□□ 2012, **7**:e35314.

48. G□mez-Pereira PR, Sch□ler M, Fuchs BM, Bennke C, Teeling H, Waldmann J, Richter M, Barbe V, Bataille E, Gl□ckner FO, Amann R: **Ge□□□ic c□□□e□□ □f □□c□□□□□ed □□□□□□□□□□□□□□□□ f□□□ c□□□□a□i□g □cea□ic □□□□i□ce□ i□ □he N□□h A□a□□ic Ocea□.** *E□□□□□□□□ □ □□□□□□□□□* 2012, **14**:52-66.

49. Hanson SR, Best MD, Wong C-HH-HH: **S□fa□a□e□: □□□□c□□□e, □echa□i□□, bi□□□gica□ac□i□i□□, i□hibi□i□□, a□d □□□□he□ic □i□i□□.** *A□□□□□ C□□□ I□□E□ E□□□* 2004, **43**:5736-5763.

50. Antelmann H, Williams RC, Miethke M, Wipat A, Albrecht D, Harwood CR, Hecker M: **The e□□□ace□□□a□a□d c□□□□□a□□ ic □□□e□□e□ □f □he □□□-i□□e□□□□□□□□□□ □□□□□□□□□□□ □□□ai□ UM23C1-2.** □□□□□□□□□□ □□□ 2005, **5**:3684-3695.

51. Long M, Ruan L, Yu Z, Xu X: **Ge□□□e □e□□e□ce □f □□□□□□□□ □□□□□ □□. S□□ai□ S9, a□ e□□□ace□□□a□ a□□□□□□□fa□a□e-□□d□ci□g bac□e□i□□ i□□□a□ed f□□□ Ma□g□□□e S□i.** *J B□□□□□□□□* 2011, **193**:4041.

52. Yang JI, Chen LC, Shih YY, Hsieh C, Chen CY, Chen WM, Chen CC: **C□□□i□g a□d cha□ac□e□i□a□i□□ □f □-aga□a□e AgaYT f□□□ □□□□ □ □□□□□□□□□ □□□□□□□ □□□□□ □□□ai□ YT.** *J B□□□□□B□□□□□* 2011, **112**:225-232.

53. Reeves AR, Wang GR, Salyers AA: **Characterization of four outer membrane proteins that are essential for starch binding and utilization in Bacteroides thetaiotaomicron**. *J Bacteriol* 1997, **179**:643-649.

54. Xie G, Bruce DC, Challacombe JF, Chertkov O, Detter JC, Gilna P, Han CS, Lucas S, Misra M, Myers GL, Richardson P, Tapia R, Thayer N, Thompson LS, Brettin TS, Henrissat B, Wilson DB, McBride MJ: **Genome sequence of the cellulolytic gliding bacterium Cytophaga hutchinsonii**. *Appl Environ Microbiol* 2007, **73**:3536-3546.

55. Koropatkin NM, Martens EC, Gordon JI, Smith TJ: **Starch catabolism by a prominent human gut symbiont is directed by the recognition of amylose helices**. *Structure (Camb, Engl : 1993)* 2008, **16**:1105.

56. Schauer K, Rodionov DA, de Reuse H: **New substrates for TonB-dependent transport: do we need the 'pin' of the Trojan horse?** *Trends Biochem Sci* 2008, **33**:330-338.

57. Krewulak KD, Vogel HJ: **TonB or not TonB: is that the question?** *Biochem Cell Biol* 2011, **89**:87-97.

58. Martens EC, Koropatkin NM, Smith TJ, Gordon JI: **Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm**. *J Biol Chem* 2009, **284**:24673-24677.

59. Shipman JA, Berleman JE, Salyers AA: **Characterization of four outer membrane proteins involved in binding starch to the cell surface of Bacteroides thetaiotaomicron**. *J Bacteriol* 2000, **182**:5365-5372.

60. Cho KH, Salyers AA: **Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by Bacteroides thetaiotaomicron**. *J Bacteriol* 2001, **183**:7224-7230.

61. González JM, Fernández-Gómez B, Fernández-Guerra A, Gómez-Consarnau L, Sánchez O, Coll-Lladó M, Del Campo J, Escudero L, Rodrbackslash'iguez-Martbackslash'inez R, Alonso-Sáez L, Latasa M, Paulsen I, Nedashkovskaya O, Lekunberri I, Pinhassi J, Pedrós-Alió C: **Genome analysis of the proteorhodopsin-containing marine bacterium Polaribacter sp. MED152 (Flavobacteria)**. *Proc Natl Acad Sci U A* 2008, **105**:8724-8729.

62. Davidson AL, Dassa E, Orelle C, Chen J: **Structure, function, and evolution of bacterial ATP-binding cassette systems**. *Microbiol Mol Biol Rev* 2008, **72**:317-64, table of contents.

63. Preiss J: **Bacterial glycogen synthesis and its regulation.** *Annu Rev Microbiol* 1984, **38**:419-458.

64. Quilès F, Polyakov P, Humbert F, Francius G: **Production of Extracellular Glycogen by *Pseudomonas fluorescens*: Spectroscopic Evidence and Conformational Analysis by Biomolecular Recognition.** *Biomacromolecules* 2012, **13**:2118-2127.

65. Ugalde JE, Parodi AJ, Ugalde RA: **De novo synthesis of bacterial glycogen: *Agrobacterium tumefaciens* glycogen synthase is involved in glucan initiation and elongation.** *Proc Natl Acad Sci U S A* 2003, **100**:10659-10663.

66. Stam MR, Danchin EGJ, Rancurel C, Coutinho PM, Henrissat B: **Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins.** *Protein Eng Des Sel* 2006, **19**:555-562.

67. Larsbrink J, Izumi A, Ibatullin FM, Nakhai A, Gilbert HJ, Davies GJ, Brumer H: **Structural and enzymatic characterization of a glycoside hydrolase family 31 α-xylosidase from *Cellvibrio japonicus* involved in xyloglucan saccharification.** *Biochem J* 2011, **436**:567-580.

68. Tan K, Tesar C, Wilton R, Keigher L, Babnigg G, Joachimiak A: **Novel α-glucosidase from human gut microbiome: substrate specificities and their switch.** *FASEB J* 2010, **24**:3939-3949.

69. Nakai H, Tanizawa S, Ito T, Kamiya K, Kim Y-MM, Yamamoto T, Matsubara K, Sakai M, Sato H, Imbe T, Okuyama M, Mori H, Sano Y, Chiba S, Kimura A: **Function-unknown glycoside hydrolase family 31 proteins, mRNAs of which were expressed in rice ripening and germinating stages, are alpha-glucosidase and alpha-xylosidase.** *J Biochem* 2007, **142**:491-500.

70. Bauer M, Kube M, Teeling H, Richter M, Lombardot T, Allers E, Würdemann CA, Quast C, Kuhl H, Knaust F, Woebken D, Bischof K, Mussmann M, Choudhuri JV, Meyer F, Reinhardt R, Amann RI, Glöckner FO: **Whole genome analysis of the marine *Bacteroidetes* '*Gramella forsetii*' reveals adaptations to degradation of polymeric organic matter.** *Environ Microbiol* 2006, **8**:2201-2213.

71. Zona R, Chang-Pi-Hin F, O'Donohue MJ, Janecek S: **Bioinformatics of the glycoside hydrolase family 57 and identification of catalytic residues in amylopullulanase from *Thermococcus hydrothermalis*.** *Eur J Biochem* 2004, **271**:2863-2872.

72. Janeček BS, Svensson B, MacGregor EA: **Structural and evolutionary aspects of two families of non-catalytic domains present in starch and glycogen binding proteins from microbes, plants and animals.** *Enzyme Microb Technol* 2011, **49**:429-440.

73. Blesák K, Janeček S: **Sequence fingerprints of enzyme specificities from the glycoside hydrolase family GH57.** *Extremophiles* 2012, **16**:497-506.

74. Bertoldo C, Antranikian G: **Starch-hydrolyzing enzymes from thermophilic archaea and bacteria.** *Curr Opin Chem Biol* 2002, **6**:151-160.

75. Seibold GM, Eikmanns BJ: **The glgX gene product of *Corynebacterium glutamicum* is required for glycogen degradation and for fast adaptation to hyperosmotic stress.** *Microbiology* 2007, **153**:2212-2220.

76. Hostinová E, Solovicová A, Gasperbackslash'ik J: **Cloning and expression of a gene for an alpha-glucosidase from *Saccharomycopsis fibuligera* homologous to family GH31 of yeast glucoamylases.** *Appl Microbiol Biotechnol* 2005, **69**:51-56.

77. McBride MJ, Xie G, Martens EC, Lapidus A, Henrissat B, Rhodes RG, Goltsman E, Wang W, Xu J, Hunnicutt DW, Staroscik AM, Hoover TR, Cheng Y-QQ-QQ, Stein JL: **Novel features of the polysaccharide-digesting gliding bacterium *Flavobacterium johnsoniae* as revealed by genome sequence analysis.** *Appl Environ Microbiol* 2009, **75**:6864-6875.

78. Pati A, Abt B, Teshima H, Nolan M, Lapidus A, Lucas S, Hammon N, Deshpande S, Cheng J-FF-FF, Tapia R, Han C, Goodwin L, Pitluck S, Liolios K, Pagani I, Mavromatis K, Ovchinikova G, Chen A, Palaniappan K, Land M, Hauser L, Jeffries CD, Detter JC, Brambilla E-MM-MM, Kannan KP, Rohde M, Spring S, Göker M, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H-PP-PP, Ivanova N: **Complete genome sequence of *Cellulophaga lytica* type strain (LIM-21).** *Stand Genomic Sci* 2011, **4**:221-232.

79. Sack EL, van der Wielen PW, van der Kooij D: ***Flavobacterium johnsoniae* as a model organism for characterizing biopolymer utilization in oligotrophic freshwater environments.** *Appl Environ Microbiol* 2011, **77**:6931-6938.

80. Franco Cairo JPL, Leonardo FC, Alvarez TM, Ribeiro DA, Büchli F, Costa-Leonardo AM, Carazzolle MF, Costa FF, Paes Leme AF, Pereira GA, Squina FM: **Functional characterization and target discovery of glycoside hydrolases from the digestome of the lower termite *Coptotermes gestroi*.** *Biotechnol Biofuels* 2011, **4**:50.

81. Pope PB, Denman SE, Jones M, Tringe SG, Barry K, Malfatti SA, McHardy AC, Cheng J-FF-FF, Hugenholtz P, McSweeney CS, Morrison M: [illegible] 2010, **107**:14793-14798.

82. Waterkeyn L, Bienfait A: [illegible] 1987, **74**:198-226.

83. Warren RAJ: [illegible] 1996, **50**:183-212.

84. Nedashkovskaya OI, Kim SB, Vancanneyt M, Snauwaert C, Lysenko AM, Rohde M, Frolova GM, Zhukova NV, Mikhailov VV, Bae KS, Oh HW, Swings J: [illegible] 2006, **56**:161-167.

85. Ivanova EP, Alexeeva YV, Flavier S, Wright JP, Zhukova NV, Gorshkova NM, Mikhailov VV, Nicolau DV, Christen R: [illegible] 2004, **54**:705-711.

86. Romanenko LA, Schumann P, Rohde M, Mikhailov VV, Stackebrandt E: [illegible] 2004, **54**:669-673.

87. Pinhassi J, Pujalte MJ, Macián MC, Lekunberri I, González JM, Pedrós-Alió C, Arahal DR: [illegible] 2007, **57**:2370-2375.

88. Gosink JJ, Woese CR, Staley JT: [illegible] 1998, **48 Pt 1**:223-235.

89. Giovannoni SJ, Bibbs L, Cho JC, Stapels MD, Desiderio R, Vergin KL, Rappé MS, Laney S, Wilhelm LJ, Tripp HJ, Mathur EJ, Barofsky DF: [illegible] **SAR11.** [illegible] 2005, **438**:82-85.

90. Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ:

[illegible]. [illegible] 2005, **309**:1242-1245.

91. Swingley WD, Sadekar S, Mastrian SD, Matthies HJ, Hao J, Ramos H, Acharya CR, Conrad AL, Taylor HL, Dejesa LC, Shah MK, O'huallachain ME, Lince MT, Blankenship RE, Beatty JT, Touchman JW: **The [illegible]**. *J B[illegible]* 2007, **189**:683-690.

92. Kalhoefer D, Thole S, Voget S, Lehmann R, Liesegang H, Wollher A, Daniel R, Simon M, Brinkhoff T: **[illegible]**. *B[illegible] C G[illegible]* 2011, **12**:324.

93. Stingl U, Desiderio RA, Cho J-C, Vergin KL, Giovannoni SJ: **The SAR92 c[illegible]**. *A[illegible] E[illegible]* 2007, **73**:2290-2296.

94. Nedashkovskaya OI, Kim SB, Han SK, Rhee MS, Lysenko AM, Falsen E, Frolova GM, Mikhailov VV, Bae KS: **[illegible]**. *I[illegible] J [illegible] E[illegible]* 2004, **54**:119-123.

95. Bowman JP: **[illegible] (Z[illegible]Be[illegible] a[illegible]d U[illegible]ha[illegible] 1944) Reiche[illegible]bach 1989 a[illegible]**. *I[illegible] J [illegible] E[illegible]* 2000, **50 P[illegible]5**:1861-1868.

96. Abt B, Lu M, Misra M, Han C, Nolan M, Lucas S, Hammon N, Deshpande S, Cheng J-FF-FF, Tapia R, Goodwin L, Pitluck S, Liolios K, Pagani I, Ivanova N, Mavromatis K, Ovchinikova G, Pati A, Chen A, Palaniappan K, Land M, Hauser L, Chang Y-JJ-JJ, Jeffries CD, Detter JC, Brambilla E, Rohde M, Tindall BJ, G[illegible]ker M, Woyke T, Bristow J, Eisen JA, Markowitz V, Hugenholtz P, Kyrpides NC, Klenk H-PP-PP, Lapidus A: **C[illegible] (IC166).** *[illegible] G[illegible]* 2011, **4**:72-80.

97. Meyer F, Paarmann D, D'Souza M, Olson R, Glass EM, Kubal M, Paczian T, Rodriguez A, Stevens R, Wilke A, Wilkening J, Edwards RA: **The [illegible] RAST [illegible]**. *B[illegible] C B[illegible]* 2008, **9**:386.

98. Yin Y, Mao X, Yang J, Chen X, Mao F, Xu Y: **dbCAN: a [illegible]**. *[illegible] A[illegible]* 2012, **40**:W445-W451.

99. Widdel F, Bak F: **Gram-negative mesophilic sulfate-reducing bacteria.** In The prokaryotes Volume *4.* 2nd edition. Edited by Balows A, Trüper H, Dworkin M, Harder W, Schleifer K. New York: Springer Verlag; 1992:3352-3378.

Fig□□e Lege□d□

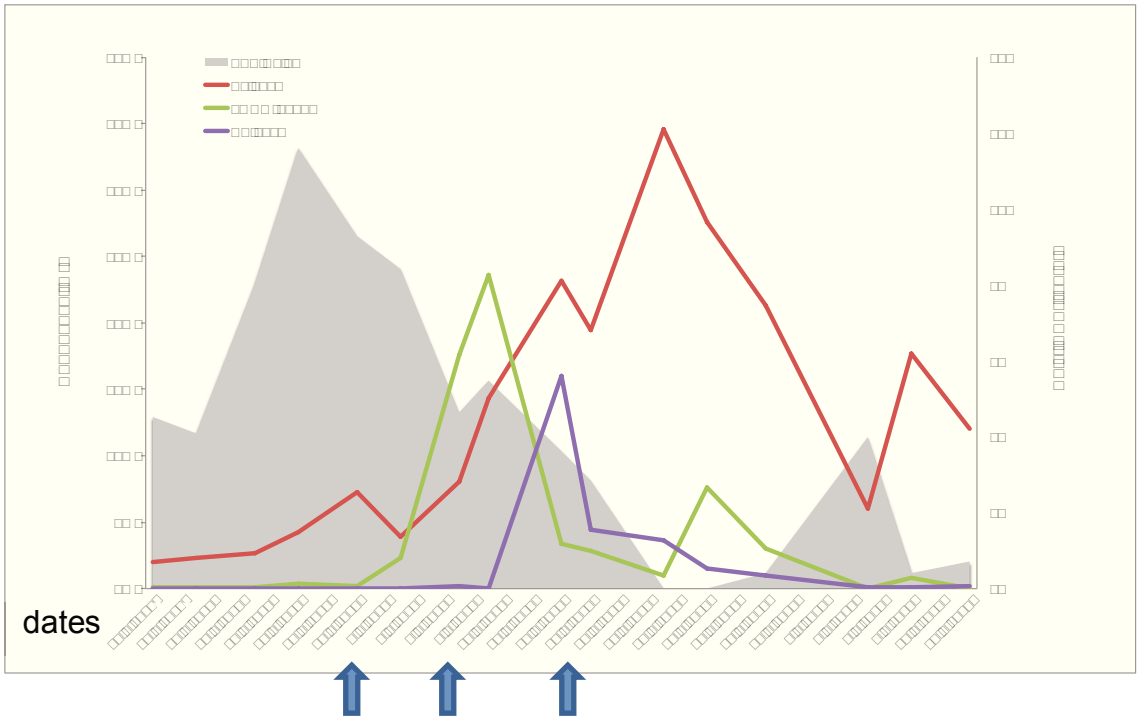


Figure 1

Chlorophyll a concentrations and abundances of the three major bacterial populations during the 2009 spring bacterioplankton bloom in the German Bight as assessed by CARD-FISH: *F□□□ □□□* spp. (probe FORM-181A), □□□□□□□□□□□□□□ spp. (probe POL740) and □□□□□□□□ spp. (probe REI731). The arrows indicate metagenome sampling dates.

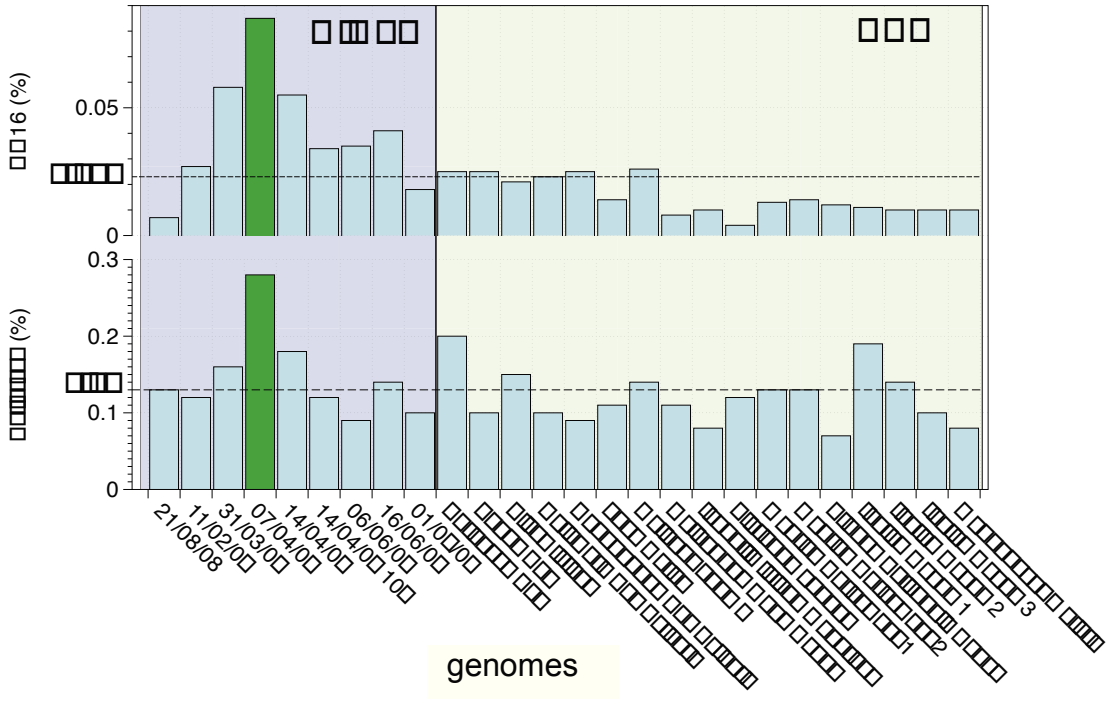


Figure 2

Gene frequencies of sulfatases and GH16 in nine MIMAS metagenomes and seventeen GOS metagenomes. The height of each column represents the gene frequency of the respective enzyme. The averages of both enzymes across all the samples are shown in dash lines.

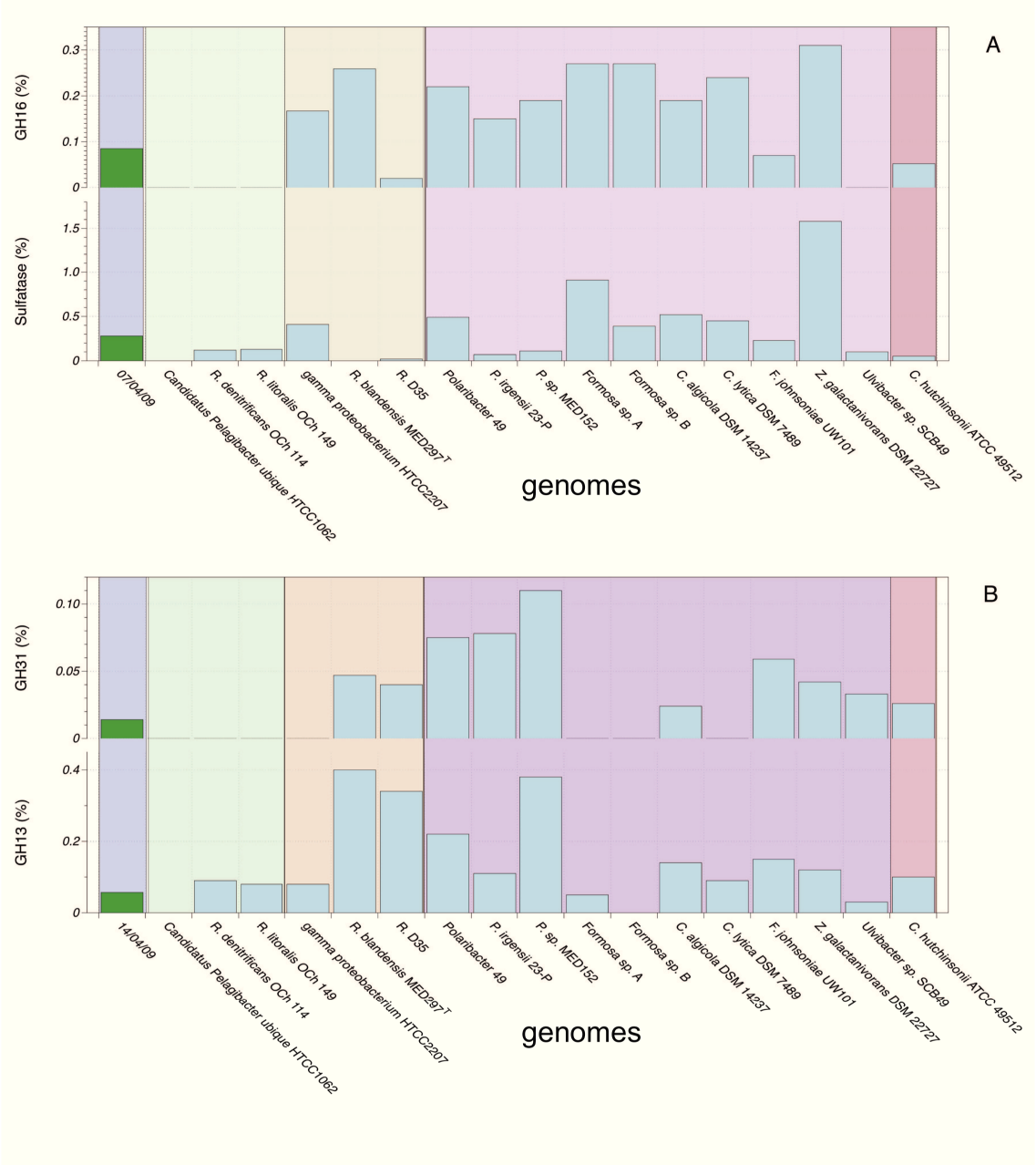


Figure 3

Comparison of GH16, sulfatases, GH13 and GH31 between two MIMAS metagenomes and nineteen reference genomes. The background colors highlight the source of the genome and the taxonomic clustering of the genomes. (blue) metagenomes from the MIMAS Project; (light green) *A*lphaproteobacteria ; (orange) *G*ammaproteobacteria ; (violet) *F*lavobacteria and together with (rosa) *B*acteroidetes

(A) Gene frequencies of sulfatases and GH16 in MIMAS metagenome on 07/04/2009 and nineteen reference bacterial genomes. The height of each column represents the gene frequency of the respective enzyme. (B) Gene frequencies of GH13 and 31 in MIMAS metagenome on 14/04/2009 and nineteen reference bacterial genomes. The height of each column represents the gene frequency of the respective enzyme.

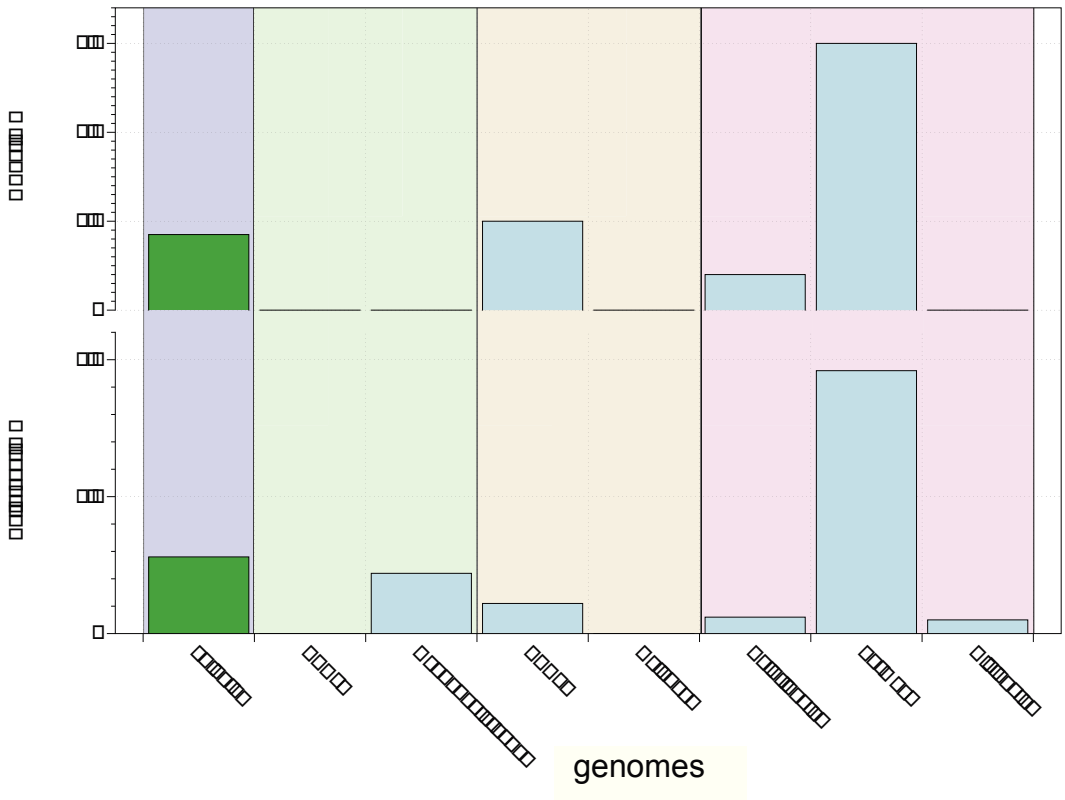


Figure 4

Gene frequencies of sulfatases and GH16 in MIMAS metagenome on 07/04/2009 and seven its taxobins. The height of each column represents the gene frequency of the respective enzyme.

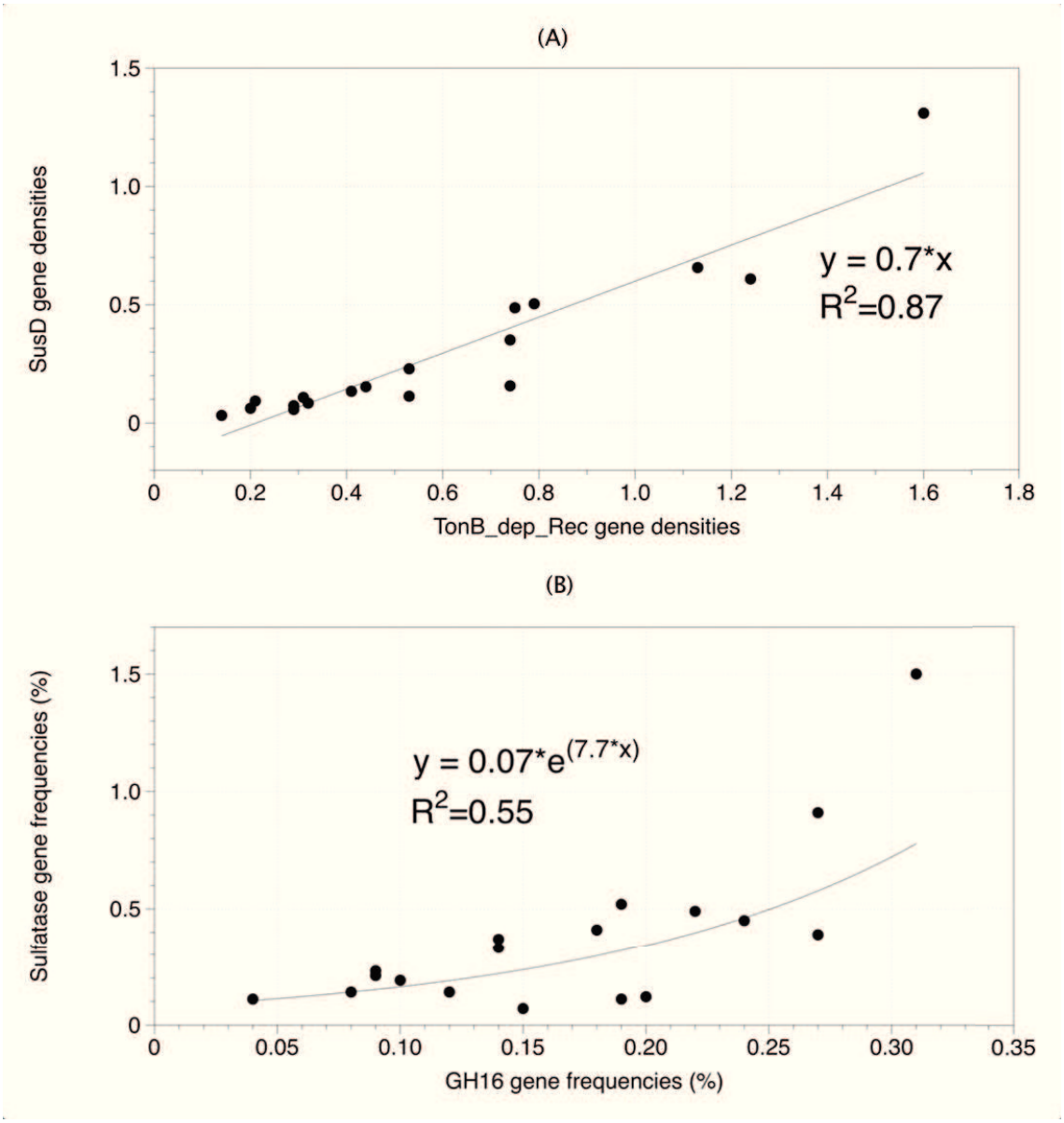


Figure 5

Correlations among the gene frequencies of SusD, TonB_dep_Rec, sulfatases and GH16 in 18 *B*□□□□□□□□□□□□taxobins and genomes (the complete list can be found in the Method section)

(A) Correlation between SusD domains and TonB_dep_Rec domains. A linear regression (y=0.7*x, R^2=0.87) was applied.

(B) Correlation between sulfatase domains and GH16. A exponential regression (y=0.07*$e^{(7.7*x)}$, R^2=0.55) was applied.

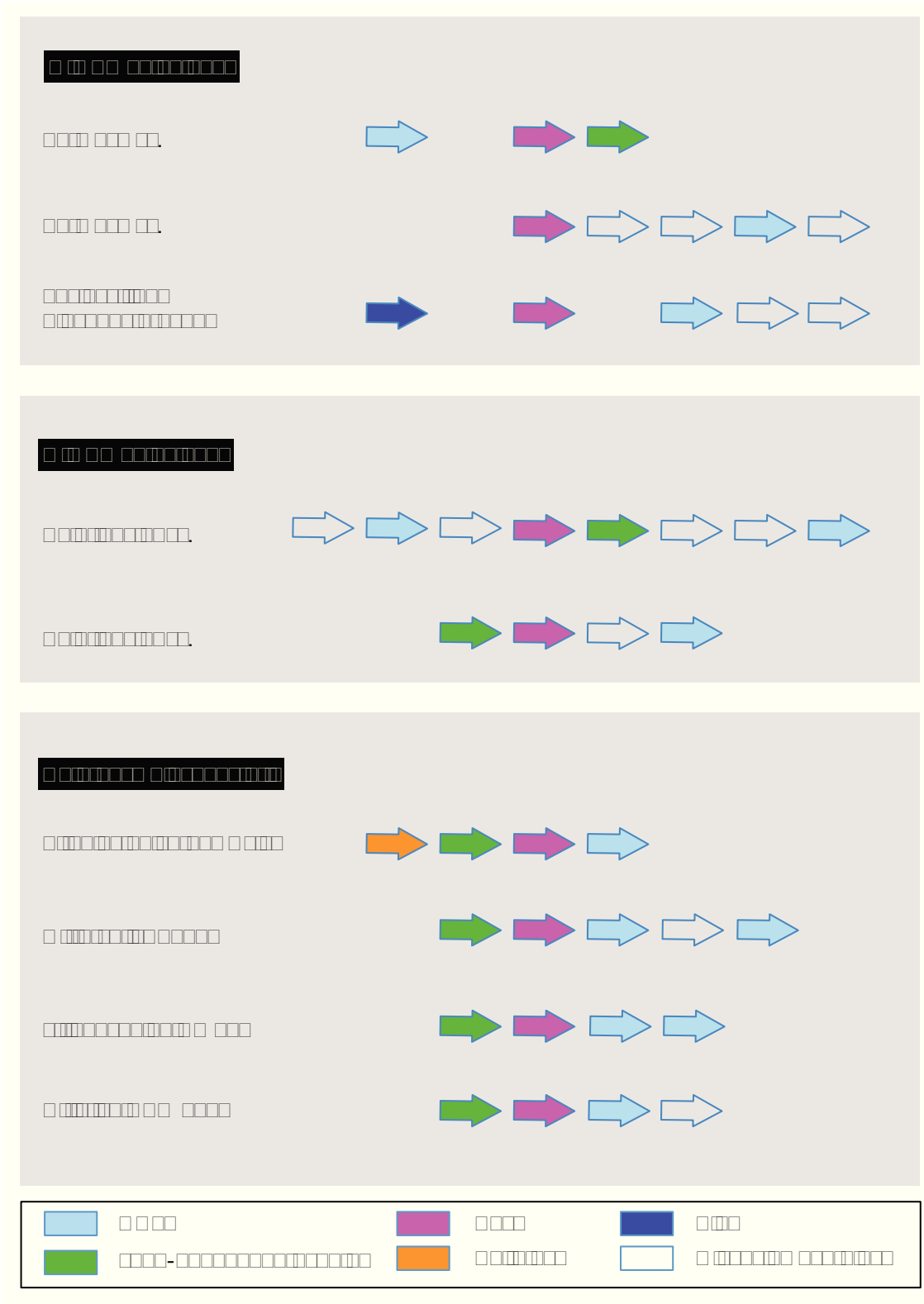
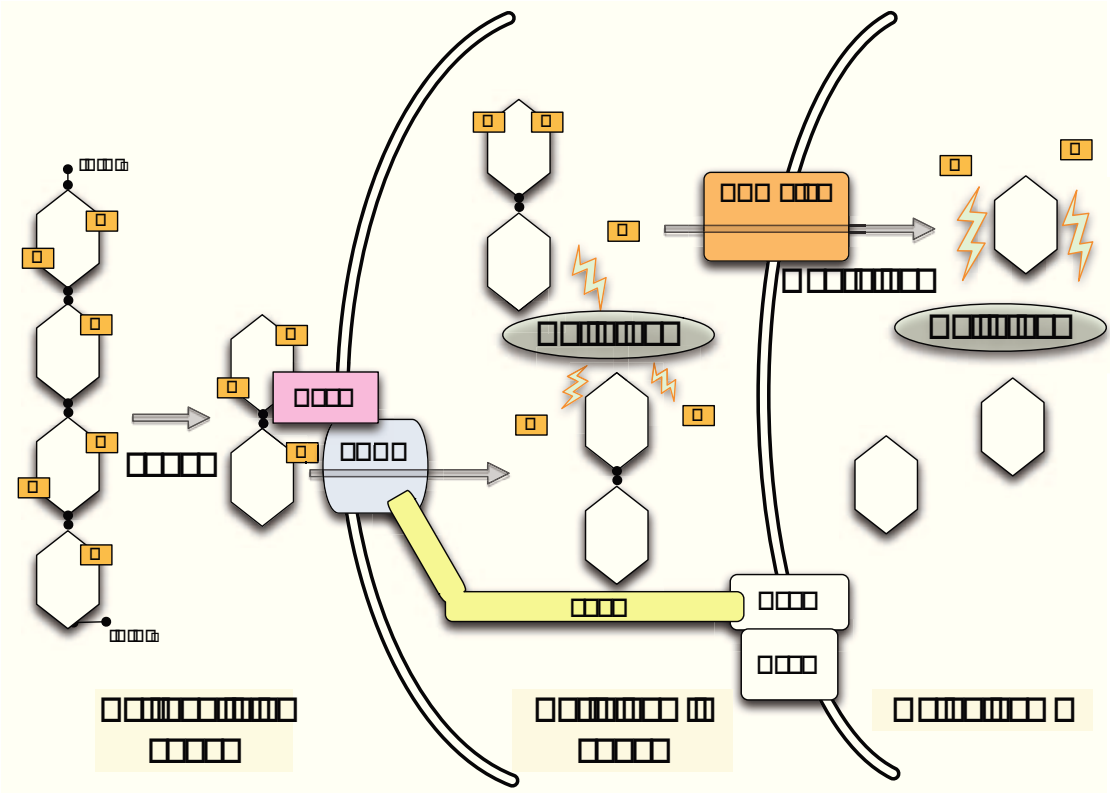


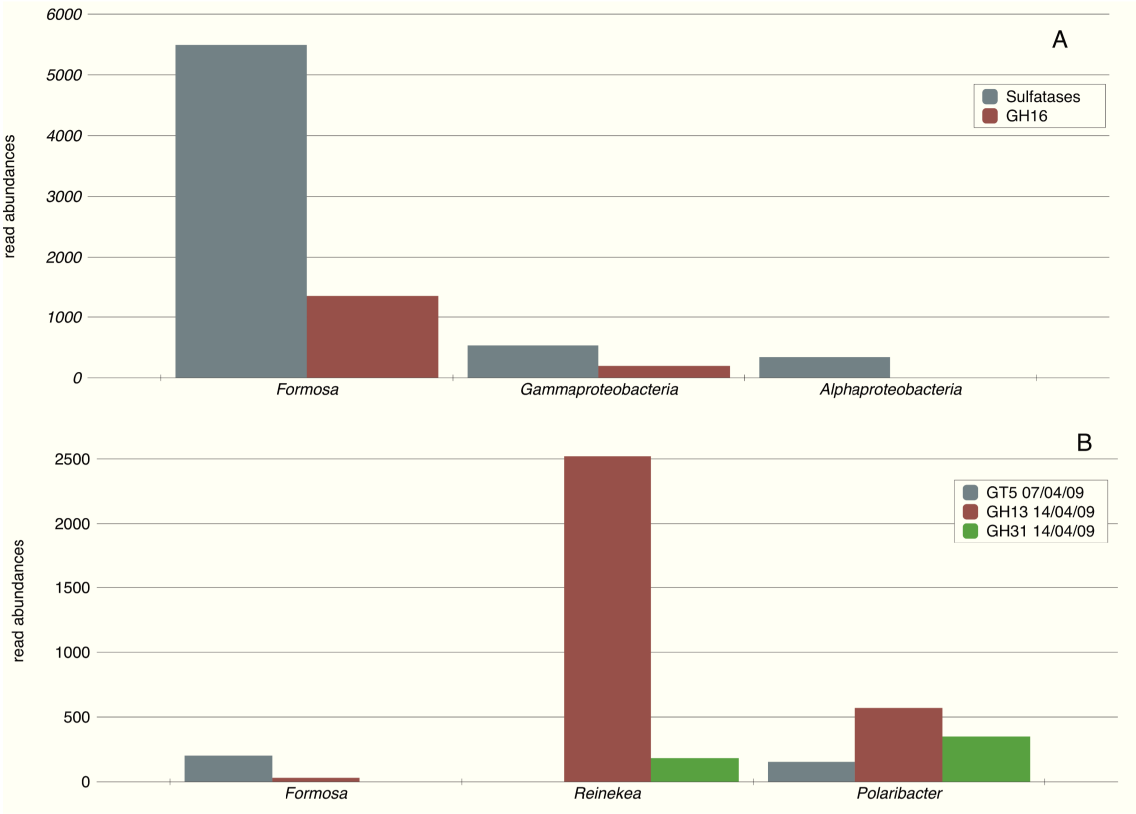
Figure 6

The gene arrangements of three flavobacterial GH16 containing-operons found in the MIMAS metagenome on 07/04/2009 and a similar operon found in [illegible] Dsij^{T} [37]. Genes annotated by CAZy are shown in blue while genes annotated by Pfam are in purple.



Supplementary 1

Hypothetical degradation model of diatom polysaccharides in *F□□□ □□□*. The degradation of the sulfated polysaccharides from diatoms involves an ordered series of steps that occur throughout the cell envelops of the *F□□□ □□□* cells. The polysaccharides are preliminarily lysed extracellularly and the shortened polysaccharides are attached to the SusD domains on the cell surface and subsequently transported through the TonB system into the periplasmic space. There periplamic sulfatases remove some of their sulfate moieties. These polysaccharides are then further imported through the ABC transporters and completely degraded intracellularly. SusD: SusD protein; TBDT: TonB-dependent transporter; TonB: TonB protein; ExbB: ExbB protein; ExbD: ExbD protein; ABC Tran: ABC transporter; S: sulfate group.



Supplementary Figure 2

(a) Read counts of sulfatases and GH16 of three bacterial taxobins on 07/04/2009. The height of each column shows the abundance of the reads that were mapped onto the respective genes in each taxobin (*F*ormosa , *G*ammaproteobacteria and *A*lphaproteobacteria).

(b) Read counts of GT5 (from 07/04/2009), GH13 and GH31 (both from 14/04/2009) of three bacterial taxobins. The height of each column shows the abundance of the reads that were mapped onto the respective genes in each taxobin (*F*ormosa , Reinekea and Polaribacter).

Supplementary Table 1: Basic statistics of the sequenced draft genomes in this study.

	F□□□ □□□ group A	*F*□□□ □□□ group B	Rei □ekea sp. D35	Pola □bacter sp. 49
contig	77	61	79	10
base pairs	2025184	2727763	3713075	2997602
ORF	1848	2543	4879	2634
CAZy modules	125	127	122	88
GH	42	44	44	28
GT	41	52	45	35

Supplementary Table 2: Expression of GH16 and GT5 of two *F*ormosa spp. isolates. Metaproteome data from the MIMAS study was re-mapped onto the draft genomes generated in this study.

	*F*ormosa group A	*F*ormosa group B
GH16	-	+
Sulfatases	-	-
GT5	-	+

Supplementary Table 3: Results of pure culture experiments growth experiments with glycogen or laminarin as supplemented carbon source. The + sign indicates subsequent increase of cell densities beyond 10^6 cells/mL is confirmed, while the - sign negates any growth.

Supplemented carbon	*F*ormosa group A	*F*ormosa group B	Rei □ekea sp. D35	Po□ari □acter sp. 49
glycogen	-	-	+	+
laminarin	+	+	-	+

2.4 Metagenomics reveals niche-differentiation and habitat-specific adaptation in surface sediments of the Logatchev hydrothermal field

Authors
Hikaru Suenaga[1,2†], Regina Schauer[1,3†], Sixing Huang[1†], Hanno Teeling[1], Chia-I Huang[1], Sven Klages[4], Richard Reinhardt[4], Frank Oliver Glöckner[1,5], Rudolf Amann[1], Anke Meyerdierks[1*]

[1] Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany
[2] National Institute of Advanced Industrial Science and Technology (AIST), Central 6, 1-1-1 Higashi, Tsukuba 305-8566, Japan
[3] Center for Geomicrobiology, Dept. of Biosciences, Ny Munkegade 116, 8000 Aarhus C, Denmark
[4] Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, D-14195 Berlin, Germany
[5] Jacobs University Bremen GmbH, Campus Ring 1, 28759 Bremen, Germany

[†] These authors contributed equally to this work

*Corresponding author: Anke Meyerdierks

Publication status
In preparation

My contribution
I assisted in the assembly of metagenomic reads. Afterwards, I was responsible for the taxonomic classification and functional annotation of the sequences. I assisted in reconstructing the metabolic pathways for carbon, nitrogen and sulfur recycling in *G*amma-, *D*elta- and *E*psilonproteobacteria. I studied the abundant proteobacterial taxobins such as Thioalkalivibrio spp., *G*eobacter spp. and Sulfurovum spp.. I compared the CAZyme profiles of the Logatchev hydrothermal field's metagenome with those from the Hot Lake samples. I reviewed the literature for β-glucan metabolism and protein glycosylation in the deep-sea. Based on these results, I proposed that the deep-sea microbial community shares some common features with surface water communities.

EM - Research Article

Metagenomics reveals niche-differentiation and habitat-specific adaptation in surface sediments of the Logatchev hydrothermal field

Hikaru Suenaga[1,2†], Regina Schauer[1,3†], Sixing Huang[1†], Hanno Teeling[1], Chia-I Huang[1], Sven Klages[4], Richard Reinhardt[4,5], Frank Oliver Glöckner[1,6], Rudolf Amann[1], Anke Meyerdierks[1]*

[1] Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany

[2] National Institute of Advanced Industrial Science and Technology (AIST), Central 6, 1-1-1 Higashi, Tsukuba 305-8566, Japan

[3] Center for Geomicrobiology, Dept. of Biosciences, Ny Munkegade 116, 8000 Aarhus C, Denmark

[4] Max Planck Institute for Molecular Genetics, Ihnestraße 63-73, D-14195 Berlin, Germany

[5] Max Planck Institute for Plant Breeding Research, Carl-von-Linné-Weg 10, D-50829 Cologne, Germany

[6] Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen, Germany

[†] These authors contributed equally to this work

*Corresponding author: Anke Meyerdierks

Running title: Microbial adaptations in Logatchev vent field surface sediments

Email addresses and telephone numbers of all authors:

Hikaru Suenaga	hsuenaga@mpi-bremen.de	+49 421 2028 545
Regina Schauer	rschauer@biology.au.dk	+45 87156506
Sixing Huang	shuang@mpi-bremen.de	+49 421 2028 928
Hanno Teeling	hteeling@mpi-bremen.de	+49 421 2028 976
Chia-I Huang	jhuang@mpi-bremen.de	+49 421 2028 905
Sven Klages	klages@molgen.mpg.de	+49 30 8413 1177
Richard Reinhardt	reinhardt@mpipz.mpg.de	+49 221 5062 810
Frank Oliver Gl□ckner	fog@mpi-bremen.de	+49 421 2028 970
Rudolf Amann	ramann@mpi-bremen.de	+49 421 2028 930
Anke Meyerdierks	ameyerdi@mpi-bremen.de	+49 421 2028 941

S□□ □a□□

The ultramafic-hosted Logatchev hydrothermal field located on the Mid Atlantic Ridge includes highly sedimented areas, characterized by conductive heating. In a previous biogeochemical study sulphur was proposed to be the driving force for microbial biomass production in the surface layer of these sediments. Here, we investigated the metagenome of the surface sediment layer in order to determine the ecology of dominant taxa and to assess genomic adaptations towards life at deep-sea hydrothermal vents. Partitioning of the metagenome into taxonomically coherent sequence bins indicated that *G□□ □ □□□□□□□□□□□□□□□□□□□□* as well as *E□□□□□□□□□□□□□□□□□□□□* oxidize reduced sulphur compounds by the Sox-dependent and -independent pathways. Both taxa seem to employ nitrate and oxygen as terminal electron acceptors, and genes for denitrification were found for both taxa. Moreover, the *E□□□□□□□□□□□□□□□□□□□□* harboured genes for nitrogen fixation, suggesting a crucial role for the sediment's nitrogen balance. The *D□□□□□□□□□□□□□□□□□□□□* contained two distinct groups of complete and incomplete -oxidizing sulphate-reducing bacteria. Analysis of carbohydrate-active enzymes indicated relevance of □□glucan turnover and glycosylation reactions for life in the sediment surface layer. A comparison with metagenomes of biofilms from hydrothermal chimney structures from two other deep-sea vent sites revealed a lower occurrence of cell motility genes in the Logatchev sediment microbial communities along with a higher occurrence of lateral gene transfer as indicated by higher levels of phage, conjugation, integron and group II intron-related genes. It is therefore proposed that adaptation strategies of microbial populations in deep-sea hydrothermal sediments are distinctly different from those in other hydrothermal habitats.

I□□□□d□c□i□□

Deep-sea hydrothermal vents are highly productive ecosystems that may account for 25% of the total imported carbon flow into the deep-sea (Maruyama □□ □□., 1998). Ecosystems at these sites are largely self-contained with chemolithoautotrophic microorganisms mediating the transfer of energy from the geothermal source in the form of biomass to organisms on higher trophic levels (Campbell □□□□., 2006; Martin □□ □□., 2008). Primary production relies almost entirely on the oxidation of reduced inorganic substrates from hydrothermal fluids such as H_2, HS^-, Fe^{2+}, and CH_4, (Nakagawa and Takai, 2008). The composition of these fluids can vary considerably, depending on the geological setting and seawater-rock interactions (Schmidt □□ □□.,

2007). Thus different energy sources are available to the vent's autochthonous microbial communities in basalt- and ultramafic-hosted hydrothermal systems.

In both systems, sulfide was reported to be an important electron donor and *Epsilonproteobacteria* as well as *Gammaproteobacteria* were shown to be dominant members of the microbial community (Gundersen et al., 1992: Brazelton et al., 2006: Campbell et al., 2006: Nakagawa and Takai, 2008: Xie et al., 2011: Yamamoto and Takai, 2011). Genomic, metagenomic and biochemical analyses revealed that *Epsilonproteobacteria* are capable of using sulphur compounds both as electron donors and acceptors. They generate energy from the oxidation of reduced sulphur compounds with O_2 or nitrate as well as the oxidation of hydrogen with sulphur, nitrate or O_2. They have been described to oxidise reduced sulphur compounds with nitrate or oxygen and to oxidize hydrogen with sulphur, nitrate or oxygen as electron acceptor. In contrast, *Gammaproteobacteria* are only known for the oxidation of sulphur compounds, presumably with oxygen as predominant terminal electron acceptor. Therefore the two dominating proteobacterial classes are thought to represent different ecophysiological strategies (for review: Yamamoto and Takai, 2011).

Despite of these metabolic differences, comparative metagenome analysis of microbial biofilms from deep-sea hydrothermal vent chimney structures revealed a high proportion of transposases pointing towards intense lateral gene transfer (LGT) (Brazelton and Baross, 2009). This, together with a high proportion of genes for DNA repair, chemotaxis and flagella assembly (Xie et al., 2011) indicated specific adaptations of microbial communities to highly dynamic conditions at deep-sea hydrothermal vents.

Previous metagenomic studies of free-living microbial communities at hydrothermal vents focused on chimney-associated biofilms (Brazelton and Baross, 2009: Xie et al., 2011). Here, we address time free-living microbial communities inhabiting hydrothermal sediments by a metagenomic approach. This site has been characterized in depth by Schauer et al. Bioinformatic classification of the sequences into taxonomically coherent bins allowed the prediction of distinct metabolic roles of each of the dominating proteobacterial classes, which emcompass beyond *Epsilonproteobacteria*, and *Gammaproteobacteria* (also *Deltaproteobacteria*). We focused our annotation on aspects of the sulphur, nitrogen, and carbon cycle. In order to elucidate unique traits of the Logatchev hydrothermal field (LHF) microbial community, furthermore the LHF metagenome was compared with two other metagenomes from deep-sea hydrothermal fields. Here it was focused on genes involved in DNA repair, motility and LGT. The results revealed common features

between hydrothermal vent communities from chimneys and sediments, but also pointed to specific adaptations with respect to □

□□□□□□□□□□□□□□□□□□□□

LHF deep-sea hydrothermal vents emit fluids of up to 350□C that are rich in sulphide, hydrogen and methane (Schmidt □□ □□., 2007). The latter two originating from rock serpentinisation at high temperatures (Gundersen □□ □□., 199???). A thick sediment layer covers large areas of the LHF (Petersen □□□□., 2010). Parts of these sediments are conductively heated, likely by underlying conduits carrying hydrothermal fluids, and are covered by patchy white sulphur-rich microbial mats (Schauer □□□□., 2011).

□□□

In this study, the metagenome of the sediment surface layer (0-1 cm; including sulphur-rich microbial mat) of site F was analysed, one of the sites previously investigated by Schauer et al. (2011). Altogether, about 408 Mbp of shotgun-metagenome sequence information was generated by 454 pyrosequencing (LHF-454 dataset) (Supplementary Table 1), and co-assembled with metagenomic fosmid, draft fosmid, and fosmid insert-end sequences (Supplementary text). The resulting contigs were classified into taxonomically coherent bins using a bioinformatic pipeline (contig classification). Subsequently, the unassembled 454 reads were mapped onto the assembly (read classification) in order to estimate the abundance of taxa within the metagenome. Additionally, 16S rRNA and 23S rRNA gene fragments of the LHF-454 dataset were taxonomically analysed. In order to evaluate whether the LHF-454 dataset reflects the microbial community composition at site F, the results of these three analyses were compared with 16S rRNA diversity and CARD-FISH data of the previous biogeochemical study (Schauer □□□□., 2011) (Table 1).

Based on read classification, 16S rRNA and 23S rRNA gene analysis 66-93% of the reads in the LHF-454 dataset originated from *B□□□□□*, and 4-9% from *A□□□□□*. This agrees largely with a proportion of <6% *A□□□□□* determined by catalysed reported deposition fluorescence □□ □□□ hybridisation (CARD-FISH) in the previous study (Schauer □□ □□. 2011). Within the *B□□□□□*, metagenome analysis revealed a predominance of □□□□□□□□□□□□□ (48-81%). Other abundant phyla were *C□□□□□□□□*(4-9%), *B□□□□□□□□□□* (2-5%), *D□□□□□□□□□□□□* (0.1-5%) and *F□□ □□□□* (1-4%). The only obvious difference to previously generated data on the microbial community composition in the sediment surface layer was the low representation of *B□□□□□□□□□□*, where previously 14-19% were determined by CARD-FISH and 19% by 16S rRNA clone library analysis. Although 16S rRNA clone library analysis cannot be considered

to be quantitative for several reasons (Fuchs und Amann 2008, Supplementary text), and the *B* probe CF319a probe used for FISH by Schauer (2011) has its limitations, it is assumed that *B* are underrepresented in the LHF metagenome.

Based on metagenome classification, the most abundant classes within the were *G* (25-43%), *E* (2-25%) and *D* (7-11%). Compared to previous FISH analyses, *G* are more frequent in the metagenome whereas *D* were lower. Again it must remain unclear whether this is due to intrinsic problems of DNA extraction based or probe-based methods (for further discussion see Supplementary text). A similar discrepancy was observed for two gammaproteobacterial families of purple sulphur bacteria, the *C* (1-13%) and *E* (0-13%). Some bioinformatic analyses tools used for taxonomic classification pointed towards significant abundances of these families, whereas both families were not represented in the previously published 16S rRNA clone libraries (Schauer , 2011). There is however consistent information for a high abundance of *D* and *H*. Therefore, we analysed the taxonomically classified sequences of the three dominant classes, *G*, *E* and *D* for their genetic capabilities with respect to the metabolism of carbon, sulphur, and nitrogen.

A

External influx of organic material into deep-sea hydrothermal fields is generally low and thus food chains are considered to be based on chemolithoautotrophic processes (Tarasov, 2006). In the LHF metagenome, indications for the presence of three of currently six known carbon fixation pathways were found: i) the Calvin-Benson-Bassham (CBB) cycle, ii) reductive tricarboxylic acid (rTCA) cycle and iii) reductive acetyl-CoA (rAcetyl-CoA) pathway (Hügler and Sievert, 2011).

Within the *G* bin, a complete set of CBB cycle genes was detected (Supplementary Table 2), with nine reads matching the key enzyme *1,5-* (RuBisCO, large or small subunit). According to homology analysis, seven reads belonged to RuBisCo form I, and two to form II. Form I and form II RuBisCOs have different CO_2 affinities; form II is preferentially used in niches with high CO_2 and low O_2, and form I in niches with low CO_2 and high O_2 partial pressures (Tabita , 2007), reflecting the strong oxygen gradients reported by Schauer et al. (2011). Both types of RuBisCo are known to coexist in the deep-sea

vent sulphur-oxidizing gammaproteobacterium [illegible] XCL-2, which enables it to survive in its aerobically complex hydrothermal vent environment (Scott [illegible], 2006). The CBB cycle is common in *G*[illegible] (H[illegible]gler and Sievert, 2011). The [illegible] gene encoding RuBisCO form I was previously detected in those gammaproteobacterial families most abundant in the LHF metagenome, *E*[illegible] (Tourova [illegible], 2010), *C*[illegible] and [illegible]. Also, the [illegible]M gene encoding RuBisCO form II is known to be present in *A*[illegible] (Kato [illegible], 2012), we also detected at site F. Therefore, *G*[illegible] at site F might have adapted to fluctuating oxygen concentrations within the upper sediment layer by harbouring two forms of RubisCO. Our still limited information is consistent with one group being adapted to .. and the second [illegible]

The epsilonproteobacterial taxonomic bin harboured all three key genes of the rTCA cycle: *A*[illegible], *2*-[illegible] and [illegible] (Supplementary Table 2). This carbon fixation pathway is common in *E*[illegible]. Its identification agrees with findings in various chemoautotrophic *E*[illegible] from deep-sea vent environments (Takai [illegible], 2005). Our taxonomic classification indicated that *H*[illegible] represented the largest family-level epsilonproteobacterial taxonomic bin in the metagenome. rTCA genes have been detected in this family before, for example in species of the genera [illegible] and [illegible] (Nakagawa [illegible], 2007: Sievert [illegible], 2008: Grote [illegible], 2012). The rTCA cycle has only been found in anaerobes and aerobes growing at low oxygen concentrations (Wahlund and Tabita, 1997), although some *E*[illegible], [illegible], may tolerate relatively high oxygen concentrations (Nakagawa [illegible], 2005).

A complete set of genes for the rAcetyl-CoA pathway was identified in the deltaproteobacterial bin (Supplementary Table 2). So far, this strictly anaerobic pathway has only been identified in chemoautotrophs, including sulphate-reducing *D*[illegible], acetogens and methanogenic *A*[illegible] (Meuer [illegible], 2002: Ragsdale, 2004). It is the dominant carbon fixation pathway in *D*[illegible] (H[illegible]gler and Sievert, 2011). The pathway has been described for members of the completely oxidising *D*[illegible], as are other species within the (Strittmatter [illegible], 2009: Brysch [illegible], 1987; Callaghan [illegible], 2012; Cravo-Laureau [illegible], 2004)). This indicates that deltaproteobacterial sulphate-reducing bacteria contribute to primary production in LHF sediments. Autotrophic sulphate-reducing bacteria are generally also capable of complete acetate oxidation (Widdel [illegible], 1983). Furthermore, two of three key genes of the rTCA cycle were identified in the deltaproteobacterial bin

(Supplementary Table 2), which indicates that some *D*[illegible] in the Logatchev sediment might also carry the rTCA cycle (H[illegible]gler [illegible], 2011). We conclude that all three dominant groups contribute to carbon fixation at LHF.

[illegible]

[illegible]

Accordingly, in the *G*[illegible] bin a high proportion of reads matched key genes of the Sox multienzyme-dependent sulfur oxidation pathway. The complete Sox-dependent pathway catalyses the oxidation of sulphide and thiosulfate to sulphate. It consists of the four key proteins SoxYZ, SoxXA, SoxB, and SoxCD (for review: Sander and Dahl, 2008). In the gammaproteobacterial bin genes for [illegible]*CD* were lacking. Without SoxCD typically sulphur deposits are formed as intermediates (Zander [illegible], 2011), which agrees with the high sulphur content and the white mat on top at Site F (Schauer [illegible], 2011). One contig (c01233) assigned to *G*[illegible] contained a [illegible] gene cluster (Supplementary Figure 1). The gene organization ([illegible]*B*[illegible]*A*) and corresponding amino acid sequences showed the highest similarities (47-61%) to the [illegible] gene cluster of the betaproteobacterium [illegible] ATCC 25259 (Beller [illegible], 2006). In addition to the Sox enzyme complex, genes coding for [illegible] (Sqr) and [illegible] (FccAB), already identified in *C*[illegible], may catalyse the oxidation of sulphite to elemental sulphur.

The presence of few genes of the reverse dissimilatory sulphate reduction (rDSR) pathway indicates the capability to remobilize and further oxidise intracellular stored sulphur via intermediary sulphide (Pott and Dahl, 1998). The conversion of sulphide to sulphate may be catalysed by [illegible]-5-[illegible] ([illegible]*AB*) and [illegible] ([illegible]). The detection of the above mentioned set of genes in the gammaproteobacterial bin agrees with the genetic equipment of most studied *G*[illegible], and of the largest gammaproteobacterial family-level bins, *C*[illegible] *E*[illegible] (Sander and Dahl, 2008). These results indicate that gammaproteobacterial strains play a crucial role for the oxidation of reduced sulphur compounds at Site F and that they may adapt to changing environmental conditions due to functional redundancy. To which extent they are responsible for sulphur deposit formation or oxidation at Site F remains to be

determined. Previous studies of hydrothermal fluids at Logatchev at least indicated that Gammaproteobacteria rather use the rDSR pathway than the Sox-pathway is used (Hugler et al , 2010). Our read frequency data on rDSR and sox genes are in line with the previous findings of

In the epsilonproteobacterial bin indications for sulphide and thiosulfate oxidation and sulphur deposit formation were found in the form of soxXB and sq -related genes (Figure 1). The higher proportion (18 reads) of sqr genes in the epsilonproteobacterial bin compared to the gammaproteobacterial bin (2 reads) suggests that mainly contribution of Epsilonproteobacteria to sulphide oxidation and sulphur accumulation. Genera within the largest epsilonproteobacterial family bin, Helicobacteraceae, are known to harbour the complete sox pathway, including soxCD (Yamamoto and Takai, 2011: Grote et al., 2012). This together with the fact that Hugler et al., (2010) could show an expression of Epsilonproteobacteria -related soxB genes also points to a direct conversion of sulphide and thiosulfate to sulphate by these organisms.

The deltaproteobacterial bin harboured predominantly sequences of families known for sulphate -reduction activity. Consequently a complete gene set for dissimilatory sulphate reduction was found, i.e. sulphate adenylyltransferase (sat), adenosine -5 -phosphosulfate reductase (aprAB) and dissimilatory sulphite reductase (dsrAB). is proposed that life of incomplete oxidising sulphate -reducing bacteria, uch as Desulfobulbaceae, which presence is indicated by the deltaproteobacterial bin (Table 1), is likely restricted to the organic rich surface sediment layer. These organoheterotrophs are able to utilize a wide variety of substrates including organic compounds (as fatty acids, alcohols, aromatic compounds and carbohydrates) as electron donors for sulphate reduction (Tor et al., 2003) and have been reported to be able to respire oxygen as protective measure (Cypionka, 2000: Lumppio et al., 2001). Facultative autotrophic complete oxidising Desulfobacteraceae are likely capable of growing in deeper sediment layers where organic matter gets rare, probably also by carbon fixation coupled to the oxidation of hydrogen (Strittmatter et al., 009) diffusing upwards from underlying conduits carrying hydrothermal fluids. This is supported by the detection of hydrogenase genes in the deltaproteobacterial taxonomic bin (data not shown).

Two key genes for sulphur metabolism, dissimilatory adenosine -5 -phosphosulfate reductase (aprA) and the SoxB component of the sulphur oxidation enzyme complex (soxB) were amplified by PCR sing environmental DNA from the sediment surface of Site F as a template. The phylogenetic distribution of detected genes was consistent with that of the metagenomic approach (Supplementary Figure 2 and 3).

At Site F oxygen got depleted a few millimeters below the sediment surface, due to high benthic oxygen consumption rates (Schauer et al. 2011). In this case, nitrate is becoming a favourable electron acceptor.

denitrifying nitrate reduction (Figure 2).

required for full nitrogenase activation (Yuvaniyama et al., 2000)

The nitrite reduction genes within the deltaproteobacterial taxobin may serve to protect sulphate-reducing bacteria from inhibitory nitrite (Kaster et al., 2007: Marietou et al., 2009), which is assumed to be predominantly produced by sulphur-oxidising nitrate reducing bacteria. It may also be used as an alternative electron acceptor, as already reported for *D* (Pfennig and Widdel, 1982), belonging to one of the abundant families detected in the dataset. However, only nitrite reduction

to ammonia is as yet know for SRB. Denitrification has as not yet been described for SRB and the function of the genes for the reduction of nitrite to nitrous oxide remains to be elucidated.

Microbial niche adaptations based on carbohydrate-active enzymes

The metagenome from Site F was analysed for carbohydrate-active enzymes (CAZymes) (Cantarel et al., 2009) (Supplementary Table 4). Among the glycoside hydrolases (GH) genes, those for α-glucan utilisation of the families GH13 (122), GH57 (37), GH77 (30) and glycosyl transferase GT35 (92) had the highest abundances. Other α-glucan-degrading families like GH31 (21) were also detected. Among the carbohydrate binding module (CBM) families CBM20 (6), CBM26 (3), CBM34 (7), CBM41 (4) and CBM48 (58) were the most prevalent. Furthermore, genes of the glycogen synthase family GT5 (174) were also abundant. Together, this CAZymes profile is indicative for active α-glucan turnover at Site F. As previously reported (Hutcheson et al., 2011), α-glucans are common organic substances in various aquatic environments where they serve as storage compounds and as connective molecules to form cell aggregates (Reina-Bueno et al., 2012). The Logatchev CAZyme profile also showed a high genetic potential for prokaryotic protein glycosylation (Supplementary Table 4), as it contained high proportions of theCAZyme families GH23 (51), GH103 (19), GH109 (61), GT41 (141), GT51 (75) and GT66 (24). Protein glycosylation is a common feature shared by both photosynthetic-based and chemosynthetic-based microbial communities (Ong et al., 1994: Moens and Vanderleyden, 1997), although the effect of protein glycosylation is only understood to a limited extent.

Habitat-specific adaptations

In order to reveal habitat specific adaptations, the LHF metagenome was compared with three other metagenomes using the MG-RAST metagenome analysis server (Reference MG-RAST). For comparison we used two metagenomes from chimney-associated microbial mats at the deep-sea hydrothermal vent fields Lost City (Brazelton and Baross, 2009) and Mothra (Xie et al., 2011), and one from filtered water of a deep-sea abyss from the North Pacific Subtropical Gyre as outgroup (DeLong et al., 2006) (Supplementary Table 5). According to the taxonomic assignment provided by MG-RAST, the microbial diversity of the hydrothermal vent field metagenomes was very similar: Proteobacteria (53-78%), Firmicutes (4-11%) and Bacteroidetes (5-7%) were the dominant phyla and Gammaproteobacteria constituted the predominant class (21-59%) (Supplementary Table 5).

All three hydrothermal vent field metagenomes were enriched in DNA repair genes

Conclusions

In this metagenome study we obtained various insights into the microbial carbon, sulphur and nitrogen cycling capabilities of the dominant microbial classes at LHF as well as their possible adaptation strategies. The combination of metagenomic information with experimental results such as microscopic observations, CARD-FISH, PCR and incubation experiments turned out to be a suitable approach for uncovering not only the metabolic potential but to extrapolate on the spatial organization of microbial taxa.

The metagenomic analysis based on the LGT pointed towards unique environmental adaptation strategies of microbial communities inhabiting hydrothermal sediments; distinctly different from those described for microbial communities associated with hydrothermal chimneys. Viruses are the most abundant biological entity and the largest source of genetic material on the planet (Suttle, 2007), and likely constitute the major vehicle for mobile elements in the ocean. Thus, the direct collection of mobile elemental DNA or RNA (metamobilome approach) will provide novel prospect on gene evolution and environmental adaptation of microbial community at marine environment.

Experimental procedures

Sampling site and sample collection

The LHF is located at the lower, eastern ridge-flank close to the axial valley at 15° N on the MAR in about 3,000 m water depth (Gebruk *et al.*, 2000). The south-eastern vent field hosts two active smokers (Quest and Irina II) as well as sediment areas covered by sulfur-mats (Anya's Garden and site F). At the times of sampling, site F was characterized by a patchy distribution of white mat areas that covered more than 50 m^2 with clumps of vent mussels scattered within the white areas (Schauer *et al.*, 2011). Sampling was conducted during the R/V Maria S. Merian cruises MSM04-3 (HYDROMAR III, 2007) and MSM10-3 (HYDROMAR IV, 2009) as described previously (Schauer *et al.*, 2011). In brief, sediment was sampled using remotely operated vehicles (Jason II, Woods Hole, USA; Kiel 6000, IFM-GEOMAR Kiel, Germany), equipped with push-corers. Aboard, the sampled cores (~8 cm diameter; 20 cm long) were sliced into 1 cm layers and directly frozen at -80 °C or directly processed for microbial turnover rate measurements as described below.

DNA extraction, and DNA sequencing

Community DNA was extracted from the LHF sample as described previously (Zhou et al., 1996). Humic substances were removed by gel purification (Plaque GP Agarose, Biozym, Hess. Oldendorf, Germany). The DNA was subsequently sequenced using the GS DNA Library Preparation Kit and a 454 GS FLX (454 Life Sciences, Branford, CT, USA). The resulting raw dataset comprised 1,152,840 reads amounting to 408 Mb.

Analysis of the 454 pyrosequencing dataset

The metagenomic 454 pyrosequencing dataset was extracted using sff_extract (http://bioinf.comav.upv.es/sff_extract/index.html), applying the -c option in order to clip adaptor regions and low quality sequence regions. Subsequently, technical replicates (Gomez-Alvarez et al. 2009) were removed applying a cut off 99% sequence identity and allowing 3 bp sequence length variation at the 5'-end of the read using a CD-Hit based (http://www.bioinformatics.org/cd-hit) replicate filter.

The resulting dataset was preprocessed (quality control and alignment) by the bioinformatics pipeline of the SILVA project (Pruesse et al., 2007). Briefly, reads shorter than 200 nucleotides and with more than 2% of ambiguities or 2% of homopolymers, respectively, were removed. Remaining reads were aligned against the SSU rRNA seed of the SILVA database release 106 (http://www.arb-silva.de/documentation/background/release-106) (Pruesse et al., 2007) using SINA Version 1.1 (Pruesse et al. 2012) whereupon non-aligned reads have not further been considered for downstream analysis. Applying this strategy, putative partial SSU rRNA gene reads within the data set could be extracted. Subsequently, remaining reads were dereplicated, clustered and classified. Dereplication (here: identification of identical reads ignoring overhangs) was done with cd-hit-est of the cd-hit package 3.1.2 (http://www.bioinformatics.org/cd-hit) using an identity criterion of 1.00 and a wordsize of 8. Remaining sequences were clustered again with cd-hit-est using an identity criterion of 0.97 (same wordsize). The longest read of each cluster was used as a reference for taxonomic classification done by a local BLAST search against the SILVA SSURef 106 NR dataset (http://www.arb-silva.de/projects/ssu-ref-nr/) using blast-2.2.22+ (http://blast.ncbi.nlm.nih.gov/Blast.cgi) with standard settings. The full SILVA taxonomic path of the best blast hit has been assigned to the reads in case the value for (% sequence identity + % alignment coverage)/2 was at least 93.0. In the final step, the taxonomic path of each cluster reference read was mapped to the additional reads within the corresponding cluster plus the corresponding replicates, identified in the

previous analysis step, to finally obtain (semi-) quantitative information (number of individual reads representing a taxonomic path).

[illegible]

Two independent assemblies were constructed from the 454 dataset, one using Newbler ver. 2.0.01.14 (454 Life Sciences), the other one using and Mira ver. 3.4 (Chevreux [illegible], 1999). Both assemblers produced different assemblies with respect to average length (Newbler: 499 bp; Mira: 536 bp), N_{50} value (Newbler: 479 bp; Mira: 504 bp) and the longest contig (Newbler: 27,079 bp; Mira: 17,938 bp). We combined both assemblies to a consensus using the Minimus2 program of the AMOS package (Sommer [illegible], 2007) generating the LHF-454 assembly. This was composed of 151,746 contigs with an average length of 551 bp, an N_{50} of 509 bp and a maximum contig size of 27,087 bp summing up to 84 Mb assembled sequence information (Supplementary Table 1).

Furthermore, 56 fosmid sequences, 42 draft fosmid sequences generated by sequencing of fosmids pools by 454 pyrosequencing, and 12 of finshed quality, (Supplementary text) as well as 6,582 fosmid insert end-sequences were assembled with the LHF-454 assembly using the Minimus2 program in order to get a better assembly for taxonomic classification of the 454 reads (LHF-Meta, Supplementary text). All analyses reported in the results section refer to sequences within the 454 dataset.

G[illegible], [illegible] [illegible]

Gene prediction and functional annotation including carbohydrate-active enzymes were carried out as reported previously (Teeling [illegible], 2012) (Supplementary text). The LHF-Meta assembly was partitioned in taxonomically coherent bins using a consensus from individual taxonomic prediction tools as described previously (Ferrer [illegible], 201/), less TaxSOM. As a result, a substantial fraction of the contigs could be classified on some taxonomic level (contig classification). Subsequently, the reads of the 454 dataset were mapped on the classified contigs in order to obtain a taxonomic classification for each individual read and to allow for quantitative analysis of the data (read classification) (superkingdom: 68%; phylum: 63%; class: 57%; order: 53%; family: 46%; genus: 47% based on the read classification).

[illegible]

Within three taxonomic bins (*G[illegible]*, *D[illegible]* [illegible] *E[illegible]*), the metabolic pathway for carbon, sulfur and nitrogen

compounds were constructed. Annotated ORFs from LHF-454 were classified by the EC number. The pathways and EC number of the enzymes were referred from the MetaCyc database (Caspi *et al.*, 2012). In order to calculate the number of the sequence reads encoding the each gene, the sequence reads were mapped on the contigs annotated by GenDB system as described in supplementary text.

Phylogenetic analysis of genes of the sulfur metabolism

In order to analyze the diversity of two key genes of the sulfur metabolisms, *aprA* (alpha subunit of adenosine phosphosulfate (APS) reductase) and *soxB* (sulfate thiohydrolase), clone libraries were constructed with environmental DNA from Site F as template. Subsequently, amplified gene fragments were phylogenetically analyzed together with genes detected in the LHF-454 dataset from Site F (Supplementary text).

Microbial mat incubations for sulfide consumption analyses

Surface sediment or the overlying sulfur mat (0.5 g) was placed in glass serum vials filled by one-third with 0.22 µm sterile-filtered bottom seawater. The vials were closed with gas-tight rubber stoppers and crimped with aluminum seals. Control vials contained either boiled sediment or solely seawater. A Na_2S stock solution was added to the seawater to a final concentration below 20 µM. All vials were incubated at 4°C on a slowly rotating table (~XXX UpM). Aliquots were taken at time intervals and fixed with zinc acetate (2%, w/v). The concentration of the resulting ZnS precipitate was measured at 663 nm by spectrophotometry according to (Trüper and Schlegel, 1964). All incubations were performed in three replicates.

Analysis of carbonhydrate-active enzymes

The annotations of carbon hydrate active enzyme from the Logatchev metagenome were carried out as described previously (Teeling *et al.*, 2012). In brief, the metagenomic sequences were subjected to a pipeline that consists of three comparison procedures: BLAST against the CAZy database (Cantarel *et al.*, 2009), Hidden Markov Models (pHMM) against Pfam (Sonnhammer *et al.*, 1997) and the dbCAN database (Yin *et al.*, 2012). The e-value cutoffs were E-30 for BLAST and E-10 for the HMM-based methods. Since the CAZy database is just a subset of the nr-database in NCBI, the best hits returned from the CAZy comparison are locally optimal but not necessarily also the globally optimal hits returned from the nr-database comparison. To detect such annotation artifacts, it is therefore meaningful to compare CAZy annotations with the results from BLAST against NCBI nr and KEGG databases.

Annotation artefacts were manually removed and statistics of eachCAZyme family were conducted.

C[illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible] [illegible]

MG-RAST (Meyer [illegible], 2008; Glass [illegible], 2010) was used to compare the LHF metagenome with three other metagenomes, namely Mothra (Xie [illegible], 2011), Lost City (Brazelton and Baross, 2009) and a North Pacific Subtropical Gyre deep abyss (DeLong [illegible], 2006) (Supplementary Table 4). Quantitative comparisons of functional genes were based on unassembled raw-reads. The dataset of Mothra was obtained from the GenBank Sequence Read Archive (SRA009990.1), and the metagenomes of Lost City and the deep abyss are publicly available on the MG-RAST servers. Genes were searched for similarity against the KEGG database with an expectation value cut-off of < E-5.

[illegible] [illegible] [illegible] [illegible]

[illegible]

[illegible] [illegible] [illegible]

We thank the captain, crew and scienti□c party of RV Meter (MSM04/3) and RV Maria S. Merian (MSM10/3) as well as the team of ROV Jason II and ROV Kiel 6000. This work was supported by grants from the priority program 1144 of the DFG. Further support came from the Center of Marine Environmental Sciences (MARUM) at the University of Bremen funded by the DFG, and the Max Planck Society.

Refe□e□ce□

Beller, H. R., Letain, T. E., Chakicherla, A., Kane, S. R., Legler, T. C., and Coleman, M. A. (2006). Whole-genome transcriptional analysis of chemolithoautotrophic thiosulfate oxidation by Thiobacillus denitrificans under aerobic versus denitrifying conditions. J Bacteriol 188: 7005-7015.

Bowles, M. W., Nigro, L. M., Teske, A. P., and Joye, S. B. (2012). Denitrification and environmental factors influencing nitrate removal in Guaymas Basin hydrothermally altered sediments. Front Microbiol 3: 377.

Brazelton, W. J., and Baross, J. A. (2009). Abundant transposases encoded by the metagenome of a hydrothermal chimney biofilm. ISME J 3: 1420-1424.

Brazelton, W. J., Schrenk, M. O., Kelley, D. S., and Baross, J. A. (2006). Methane- and sulfur-metabolizing microbial communities dominate the Lost City hydrothermal field ecosystem. Appl Environ Microbiol 72: 6257-6270.

Brunet, R. C., and Garcia‐il, L. J. (2006). Sulfide‐nduced dissimilatory nitrate reduction to ammonia in anaerobic freshwater sediments. FEMS Microbiology Ecology 21: 131-138.

Brysch, K., Schneider, C., Fuchs, G., and Widdel, F. (1987). Lithoautotrophic growth of sulfate-reducing bacteria, and description of Desulfobacterium autotrophicum gen. nov., sp. nov. Archives of microbiology 148: 264-274.

Callaghan, A. V., Morris, B. E., Pereira, I. A., McInerney, M. J., Austin, R. N., Groves, J. T., et al. (2012). The genome sequence of Desulfatibacillum alkenivorans AK-01: a blueprint for anaerobic alkane oxidation. Environ Microbiol 14: 101-113.

Campbell, B. J., Engel, A. S., Porter, M. L., and Takai, K. (2006). The versatile andepsi;-proteobacteria: key players in sulphidic habitats. Nature Reviews Microbiology 4: 458-468.

Cantarel, B. L., Coutinho, P. M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009). The Carbohydrate-Active EnZymes database (CAZy): an expert resource for glycogenomics. Nucleic Acids Res 37(suppl 1): D233-D238.

Caspi, R., Altman, T., Dreher, K., Fulcher, C. A., Subhraveti, P., Keseler, I. M., et al. (2012). The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. Nucleic Acids Res 40(Database issue): D742-53.

Chevreux, B., Wetter, T., and Suhai, S. (1999). Genome sequence assembly using trace signals and additional sequence information. Proceedings from Computer Science and Biology: Proceedings of the German Conference on Bioinformatics (GCB).

Cravo-Laureau, C., Matheron, R., Joulian, C., Cayol, J. L., and Hirschler-Rea, A. (2004). Desulfatibacillum alkenivorans sp. nov., a novel n-alkene-degrading, sulfate-reducing bacterium, and emended description of the genus Desulfatibacillum. Int J Syst Evol Microbiol 54(Pt 5): 1639-1642.

Cypionka, H. (2000). Oxygen respiration by desulfovibrio species. Annu Rev Microbiol 54: 827-848.

DeLong, E. F., Preston, C. M., Mincer, T., Rich, V., Hallam, S. J., Frigaard, N. U., et al. (2006). Community genomics among stratified microbial assemblages in the ocean's interior. Science 311: 496-503.

Ferrer, M., Werner, J., Chernikova, T. N., Bargiela, R., Fernandez, L., La Cono, V., et al. (2012). Unveiling microbial life in the new deep-sea hypersaline Lake Thetis. Part II: a metagenomic study. Environ Microbiol 14: 268-281.

Gebruk, A. V., Southward, E. C., Kennedy, H., and Southward, A. J. (2000). Food sources, behaviour, and distribution of hydrothermal vent shrimps at the Mid-Atlantic Ridge. Journal of the Marine Biological Association of the UK 80: 485-499.

Glass, E. M., Wilkening, J., Wilke, A., Antonopoulos, D., and Meyer, F. (2010). Using the metagenomics RAST server (MG-RAST) for analyzing shotgun metagenomes. Cold Spring Harbor Protocols 2010: pdb. prot5368.

Grote, J., Schott, T., Bruckner, C. G., Glockner, F. O., Jost, G., Teeling, H., et al. (2012). Genome and physiology of a model Epsilonproteobacterium responsible for sulfide detoxification in marine oxygen depletion zones. Proc Natl Acad Sci U S A, 109: 506-510.

Gundersen, J. K., Jorgensen, B. B., Larsen, E., and Jannasch, H. W. (1992). Mats of giant sulphur bacteria on deep-sea sediments due to fluctuating hydrothermal flow. Nature 360: 454-456.

Hickman, J. W., Tifrea, D. F., and Harwood, C. S. (2005). A chemosensory system that regulates biofilm formation through modulation of cyclic diguanylate levels. Proc Natl Acad Sci U S A, 102: 14422-14427.

Huber, J. A., Mark Welch, D. B., Morrison, H. G., Huse, S. M., Neal, P. R., Butterfield, D. A., and Sogin, M. L. (2007). Microbial population structures in the deep marine biosphere. Science, 318: 97-100.

H□gler, M., Gartner, A., and Imhoff, J. F. (2010). Functional genes as markers for sulfur cycling and CO2 fixation in microbial communities of hydrothermal vents of the Logatchev field. FEMS Microbiol Ecol 73: 526-537.

H□gler, M., Petersen, J. M., Dubilier, N., Imhoff, J. F., and Sievert, S. M. (2011). Pathways of carbon and energy metabolism of the epibiotic community associated

with the deep-sea hydrothermal vent shrimp Rimicaris exoculata. PLoS One 6:e16018.

H□gler, M., and Sievert, S. M. (2011). Beyond the Calvin cycle: autotrophic carbon fixation in the ocean. Annu Rev Marine Science, 3:261-289

Hutcheson, S. W., Zhang, H., and Suvorov, M. (2011). Carbohydrase systems of Saccharophagus degradans degrading marine complex polysaccharides. Mar Drugs 9: 645-665.

Kaster, K. M., Grigoriyan, A., Jennneman, G., and Voordouw, G. (2007). Effect of nitrate and nitrite on sulfide production by two thermophilic, sulfate-reducing enrichments from an oil field in the North Sea. Applied microbiology and biotechnology 75: 195-203.

Kato, S., Nakawake, M., Ohkuma, M., and Yamagishi, A. (2012). Distribution and phylogenetic diversity of cbbM genes encoding RubisCO form II in a deep-sea hydrothermal field revealed by newly designed PCR primers. Extremophiles 16: 277-283.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., and Glockner, F. O. (2012). Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. Nucleic Acids Res, in press. doi:10.1093/nar/gks808

Lumppio, H. L., Shenvi, N. V., Summers, A. O., Voordouw, G., and Kurtz, D. M. J. (2001). Rubrerythrin and rubredoxin oxidoreductase in Desulfovibrio vulgaris: a novel oxidative stress protection system. J Bacteriol, 183: 101-108.

Marietou, A., Griffiths, L., and Cole, J. (2009). Preferential reduction of the thermodynamically less favorable electron acceptor, sulfate, by a nitrate-reducing strain of the sulfate-reducing bacterium Desulfovibrio desulfuricans 27774. J Bacteriol 191: 882-889.

Martin, W., Baross, J., Kelley, D., and Russell, M. J. (2008). Hydrothermal vents and the origin of life. Nature Reviews Microbiology 6: 805-814.

Medini, D., Donati, C., Tettelin, H., Masignani, V., and Rappuoli, R. (2005). The microbial pan-genome. Curr Opin Genet Dev, 15: 589-594.

Meuer, J., Kuettner, H. C., Zhang, J. K., Hedderich, R., and Metcalf, W. W. (2002). Genetic analysis of the archaeon Methanosarcina barkeri Fusaro reveals a central role for Ech hydrogenase and ferredoxin in methanogenesis and carbon fixation. Proc Natl Acad Sci U S A, 99: 5632-5637.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al. (2008). The metagenomics RAST server□public resource for the automatic phylogenetic and functional analysis of metagenomes. BMC bioinformatics 9: 386.

Moens, S., and Vanderleyden, J. (1997). Glycoproteins in prokaryotes. Archives of microbiology 16: 169-175.

Nakagawa, S., Takai, K., Inagaki, F., Hirayama, H., Nunoura, T., Horikoshi, K., and Sako, Y. (2005). Distribution, phylogenetic diversity and physiological characteristics of epsilon-Proteobacteria in a deep-sea hydrothermal field. Environ Microbiol 7: 1619-1632.

Nakagawa, S., Takaki, Y., Shimamura, S., Reysenbach, A. L., Takai, K., and Horikoshi, K. (2007). Deep-sea vent epsilon-proteobacterial genomes provide insights into emergence of pathogens. Proc Natl Acad Sci U S A, 104: 12146-12150.

Nakagawa, S., and Takai, K. (2008). Deep-sea vent chemoautotrophs: diversity, biochemistry and ecological significance. FEMS Microbiol Ecol 65: 1-14.

Ong, E., Kilburn, D. G., Miller Jr, R. C., and Warren, R. A. (1994). Streptomyces lividans glycosylates the linker region of a beta-1, 4-glycanase from Cellulomonas fimi. Journal of bacteriology 176: 999-1008.

Petersen, J. M., Ramette, A., Lott, C., Cambon-Bonavita, M. A., Zbinden, M., and Dubilier, N. (2010). Dual symbiosis of the vent shrimp Rimicaris exoculata with filamentous gamma- and epsilonproteobacteria at four Mid-Atlantic Ridge hydrothermal vent fields. Environ Microbiol 12: 2204-2218.

Pfennig, N., and Widdel, F. (1982). The bacteria of the sulphur cycle. Philos Trans R Soc Lond B Biol Sci 298: 433-441.

Pott, A. S., and Dahl, C. (1998). Sirohaem sulfite reductase and other proteins encoded by genes at the dsr locus of Chromatium vinosum are involved in the oxidation of intracellular sulfur. Microbiology 144(Pt 7): 1881-1894.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Gl□kner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucleic acids research 35: 7188-7196.

Pruski, A. M., and Dixon, D. R. (2003). Toxic vents and DNA damage: first evidence from a naturally contaminated deep-sea environment. Aquat Toxicol 64: 1-13.

Ragsdale, S. W. (2004). Life with carbon monoxide. Critical reviews in biochemistry and molecular biology 39: 165-195.

Reina-Bueno, M., Argando□, M., Salvador, M., Rodr□uez-Moya, J., Iglesias-Guerra, F., Csonka, L. N., et al. (2012). Role of Trehalose in Salinity and Temperature Tolerance in the Model Halophilic Bacterium Chromohalobacter salexigens. PloS one, 7: e33587.

Sander, J., and Dahl, C. (2008). Metabolism of inorganic sulfur compounds in purple bacteria. Advances in Photosynthesis and Respiration 28: The Purple Phototrophic Bacteria 595-622.

Schauer, R., R□, H., Augustin, N., Gennerich, H. H., Peters, M., Wenzhoefer, F., et al. (2011). Bacterial sulfur cycling shapes microbial communities in surface sediments of an ultramafic hydrothermal vent field. Environmental microbiology 13: 2633-2648.

Schmidt, K., Koschinsky, A., Garbe-Sch□berg, D., de Carvalho, L. M., and Seifert, R. (2007). Geochemistry of hydrothermal fluids from the ultramafic-hosted Logatchev hydrothermal field, 15 N on the Mid-Atlantic Ridge: temporal and spatial investigation. Chemical geology 242: 1-21.

Schrenk, M. O., Kelley, D. S., Bolton, S. A., and Baross, J. A. (2004). Low archaeal diversity linked to subseafloor geochemical processes at the Lost City Hydrothermal Field, Mid-Atlantic Ridge. Environ Microbiol 6: 1086-1095.

Scott, K. M., Sievert, S. M., Abril, F. N., Ball, L. A., Barrett, C. J., Blake, R. A., et al. (2006). The genome of deep-sea vent chemolithoautotroph Thiomicrospira crunogena XCL-2. PLoS Biol 4: e383.

Sievert, S. M., Scott, K. M., Klotz, M. G., Chain, P. S., Hauser, L. J., Hemp, J., et al. (2008). Genome of the epsilonproteobacterial chemolithoautotroph Sulfurimonas denitrificans. Appl Environ Microbiol 74: 1145-1156.

Sommer, D., Delcher, A., Salzberg, S., and Pop, M. (2007). Minimus: a fast, lightweight genome assembler. BMC bioinformatics 8: 64.

Sonnhammer, E. L. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins Structure Function and Genetics 28: 405-420.

Sorokin, D. Y., and Kuenen, J. G. (2005). Chemolithotrophic haloalkaliphiles from soda lakes. FEMS microbiology ecology 52: 287-295.

Sorokin, D. Y., and Kuenen, J. G. (2006). Haloalkaliphilic sulfur‐xidizing bacteria in soda lakes. FEMS microbiology reviews 29: 685-702.

Sorokin, D. Y., Kuenen, J. G., and Jetten, M. S. (2001). Denitrification at extremely high pH values by the alkaliphilic, obligately chemolithoautotrophic, sulfur-oxidizing bacterium Thioalkalivibrio denitrificans strain ALJD. Arch Microbiol 175: 94-101.

Strittmatter, A. W., Liesegang, H., Rabus, R., Decker, I., Amann, J., Andres, S., et al. (2009a). Genome sequence of Desulfobacterium autotrophicum HRM2, a marine sulfate reducer oxidizing organic carbon completely to carbon dioxide. Environ Microbiol 11: 1038-1055.

Strittmatter, A. W., Liesegang, H., Rabus, R., Decker, I., Amann, J., Andres, S., et al. (2009b). Genome sequence of Desulfobacterium autotrophicum HRM2, a marine sulfate reducer oxidizing organic carbon completely to carbon dioxide. Environ Microbiol 11: 1038-1055.

Suttle, C. A. (2007). Marine viruses□ ajor players in the global ecosystem. Nature Reviews Microbiology 5: 801-812.

Tabita, F. R., Hanson, T. E., Li, H., Satagopan, S., Singh, J., and Chan, S. (2007). Function, structure, and evolution of the RubisCO-like proteins and their RubisCO homologs. Microbiol Mol Biol Rev 71: 576-599.

Takai, K., Campbell, B. J., Cary, S. C., Suzuki, M., Oida, H., Nunoura, T., et al. (2005). Enzymatic and genetic characterization of carbon and energy metabolisms by deep-sea hydrothermal chemolithoautotrophic isolates of Epsilonproteobacteria. Appl Environ Microbiol 71: 7310-7320.

Tarasov, V. G. (2006). Effects of shallow-water hydrothermal venting on biological communities of coastal marine ecosystems of the western Pacific. Adv Mar Biol 50: 267-421.

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., et al. (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science 336: 608-611.

Tor, J. M., Amend, J. P., and Lovley, D. R. (2003). Metabolism of organic compounds in anaerobic, hydrothermal sulphate-reducing marine sediments. Environ Microbiol 5: 583-591.

Tourova, T. P., Kovaleva, O. L., Sorokin, D. Y., and Muyzer, G. (2010). Ribulose-1,5-bisphosphate carboxylase/oxygenase genes as a functional marker for chemolithoautotrophic halophilic sulfur-oxidizing bacteria in hypersaline habitats. Microbiology 156(Pt 7): 2016-2025.

Tran, H. T., Krushkal, J., Antommattei, F. M., Lovley, D. R., and Weis, R. M. (2008). Comparative genomics of Geobacter chemotaxis genes reveals diverse signaling function. BMC Genomics 9: 471.

Tr□per, H. G., and Schlegel, H. G. (1964). Sulphur metabolism in Thiorhodaceae I. Quantitative measurements on growing cells of *C□□□□ □□□□□ □□□□□□* Antonie van Leeuwenhoek 30: 225-238.

Wahlund, T. M., and Tabita, F. R. (1997). The reductive tricarboxylic acid cycle of carbon dioxide assimilation: initial studies and purification of ATP-citrate lyase from the green sulfur bacterium Chlorobium tepidum. J Bacteriol 179: 4859-4867.

Widdel, F., Kohring, G. W., and Mayer, F. (1983). Studies on dissimilatory sulfate-reducing bacteria that decompose fatty acids. Archives of Microbiology 134: 286-294.

Xie, W., Wang, F., Guo, L., Chen, Z., Sievert, S. M., Meng, J., et al. (2011). Comparative metagenomics of microbial communities inhabiting deep-sea hydrothermal vent chimneys with contrasting chemistries. ISME J 5: 414-426.

Yamamoto, M., and Takai, K. (2011). Sulfur metabolisms in epsilon- and gamma-proteobacteria in deep-sea hydrothermal fields. Front Microbiol 2: 192.

Yin, Y., Mao, X., Yang, J., Chen, X., and Mao. F. , et al.(2012). dbCAN: a web resource for automated carbohydrate-active enzyme annotation. Nucleic Acids Res 40: w445-w451.

Yuvaniyama, P., Agar, J. N., Cash, V. L., Johnson, M. K., and Dean, D. R. (2000). NifS-directed assembly of a transient [2Fe-2S] cluster within the NifU protein. Proc Natl Acad Sci U S A 97: 599-604.

Zander, U., Faust, A., Klink, B. U., de Sanctis, D., Panjikar, S., Quentmeier, A.,et al. (2011). Structural basis for the oxidation of protein-bound sulfur by the sulfur cycle molybdohemo-enzyme sulfane dehydrogenase SoxCD. Journal of Biological Chemistry 286: 8349-8360.

Zhou, J., Bruns, M. A., and Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. Applied and environmental microbiology 62: 316-322.

Zopfi, J., Kjaer, T., Nielsen, L. P., and Jorgensen, B. B. (2001). Ecology of Thioploca spp.: nitrate and sulfur storage in relation to chemical microgradients and influence of Thioploca spp. on the sedimentary nitrogen cycle. Appl Environ Microbiol 67: 5530-5537.

Fig□□e□ a□d Tab□e□

Table 1 Microbial diversity and abundance at Site F.

	□ead	16S □RNA □ e□age□□□ e	23S □RNA □ e□age□□□ e	CARD-FISH	16S □RNA c□□□e □ib□a□ie□
□□□□□□□	3.6	8.8	7.3	<6.0	n.d.
□□□□□□□□	65.9	91.2	92.7	???	
□□□□□□□□□□□□	0.1	1.4	0.6	n.d.	0.9
□□□□□□□□□□□□	1.2	0.0	0.0	n.d.	0.0
□□□□□□□□□□□□	2.1	4.9	3.4	14.0-19.0	18.9
□□□□□□□□□	2.1	8.8	1.4	n.d.	4.5
□□□□□□□□□□□□	1.9	0.4	0.0	n.d.	1.8
□□□□□□□□□□□□□□	0.1	4.6	1.1	n.d.	0.9
□□□ □□□□□	3.5	0.7	0.8	n.d.	0.9
□□□□□□□ □□□□□□	0.9	1.8	2.2	n.d.	1.8
□□□□□□□□□□□□	48.3	56.3	80.9	n.d.	60.4
□□□□□□□□□□□□□□□□	4.6	0.0	0.6	n.d.	0.0
□□□□□□□□□□□□□□□	1.9	0.0	0.3	n.d.	1.8
□□□□□□□□□□□□□□□	4.0	0.0	1.1	n.d.	0.0
□ □□□□□□□□□□□	0.0	0.0	1.1	n.d.	0.0
□ □□ □ □□□□□□□□□□□□	24.7	26.4	43.0	5.0	35.1
A□□□□□□□□□□□□□	0.0	0.7	2.5	n.d.	0.0
C□□□□ □□□□□□□	1.2	4.9	12.9	n.d.	0.0
A□□□□□□ □□□□	0.2	0.0	11.0	n.d.	0.0
H□□□□□□ □□□□	0.0	2.1	0.0	n.d.	0.0
□□□□□□□□□□□	0.9	0.4	2.0	n.d.	0.0
C□□□□□□□□□□□	0.0	1.4	0.0	n.d.	0.0
E□□□□□□□□□□□□□□	12.7	0.0	13.5	n.d.	0.0
A□□□□□ □□□□	2.0	0.0	1.1	n.d.	0.0
□□□□□□□□	0.9	0.0	1.4	n.d.	0.0
□□□□□□□□□□□	8.7	0.0	10.7	n.d.	0.0
□ □□□□□□□□□□□	1.6	0.7	6.5	n.d.	0.0
□ □□□□□□□□□□□	0.1	0.0	6.5	n.d.	0.0
□□□□□□□□□□□	0.0	1.1	1.4	n.d.	0.9
□□□□□□□□	0.6	2.5	0.0	n.d.	0.9
□□□□□□□□□□□□	9.6	7.0	11.2	21.0	14.4
D□□□□□□□□□□□	0.0	0.4	1.4	n.d.	1.8
D□□□□□□□□□□□	1.8	3.5	7.0	n.d.	5.4
D□□□□□□□□□□□	1.0	0.0	5.3	n.d.	0.0
D□□□□□□□□□□□	0.2	1.1	1.7	n.d.	1.8
G□□□□□□□□□□□	1.6	0.0	0.3	n.d.	0.0
□□□□□□□□□□□	0.0	1.4	0.0	n.d.	0.0
□□□□□□□□□□□□□□	2.2	22.9	25.0	21.0	8.1

H	0.3	21.8	24.2	n.d.	8.1
	0.0	2.5	0.0	n.d.	0.0
	0.2	10.6	11.5	n.d.	5.4
	0.9	8.8	12.6	n.d.	2.7
	0.1	2.1	0.3	n.d.	0.0

The relative abundance of classified 454 reads, 16S rRNA and 23S rRNA gene fragments is given in comparision to 16S rRNA clone library and CARD-FISH studies on Site F surface sediment by Schauer et al. (2011). Listed are domains, phyla, proteobacterial classes, families within the proteobacterial classes, and genera within these most abundant families with at least 1% relative abundance in at least one of the analyses. Clades without cultured representatives have not been considered.

Table 2 Abundance of genes involved in the environmental adaptation in three metagenomes

	Logatchev[a]	Lost City[a]	Mothra[a]	Deep Abyss[a]
DNA repair				
Homologous recombination	2.87	3.42	2.85	1.12
Mismatch repair	1.45	1.23	1.42	0.992
Nucleotide excision repair	1.05	0.95	0.97	0.512
Cell motility				
Bacterial chemotaxis	0.34	0.83	1.21	0.22
Flagella assembly	0.68	2.03	2.38	0.35
Mobile element				
Transposable element	0.18	0.17	0.17	0.074
Phage	0.75	0.44	0.50	0.409
Conjugation	0.19	0.05	0.06	0.009
Integron	0.047	0.004	0.007	0.009
Group II intron	0.23	0.036	0.13	0.037

[a]The number indicates the relative abundance of reads in each metagenomic data set in percent.

The classification of gene is based on KEGG (DNA repair and Cell motility) and SEED Subsystems (Mobile element).

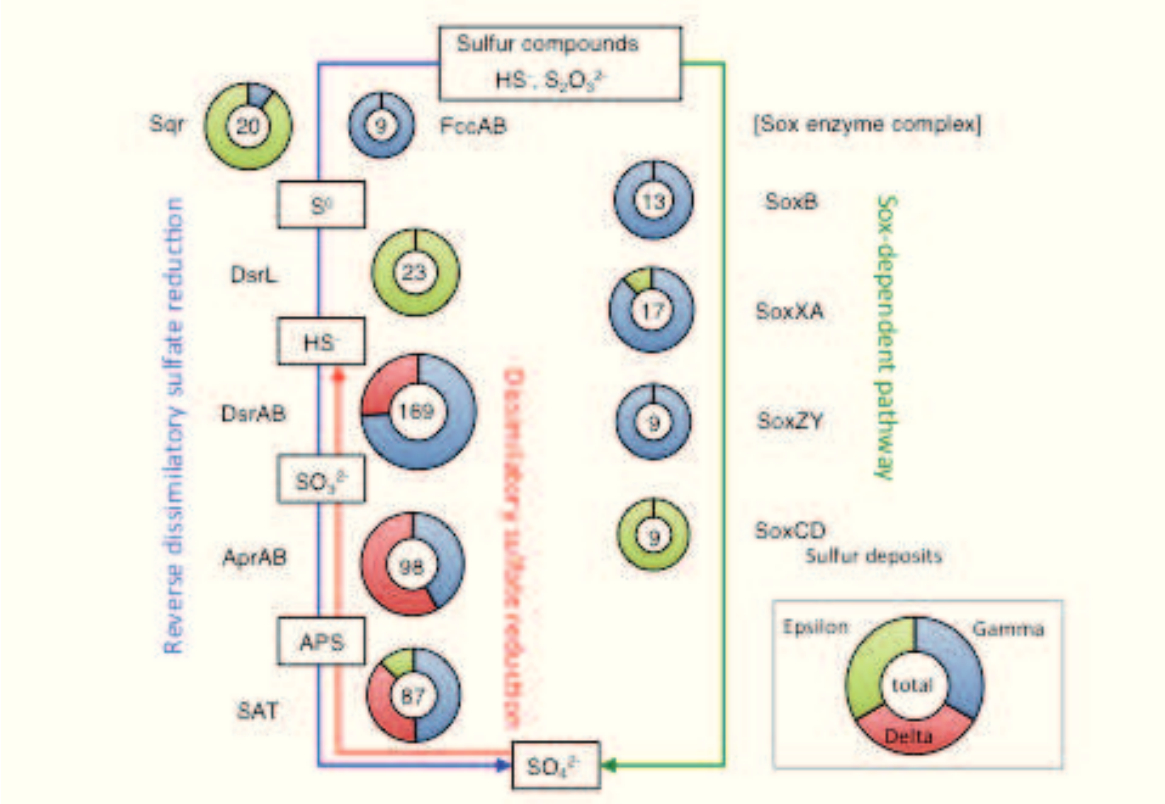


Figure 1: Components of the dissimilatory sulfur metabolism pathway identified in this study. The doughnut-like chart indicates the proportion of reads in each taxonomic bin. The number in the middle of the chart shows the sequence reads matching the corresponding gene.

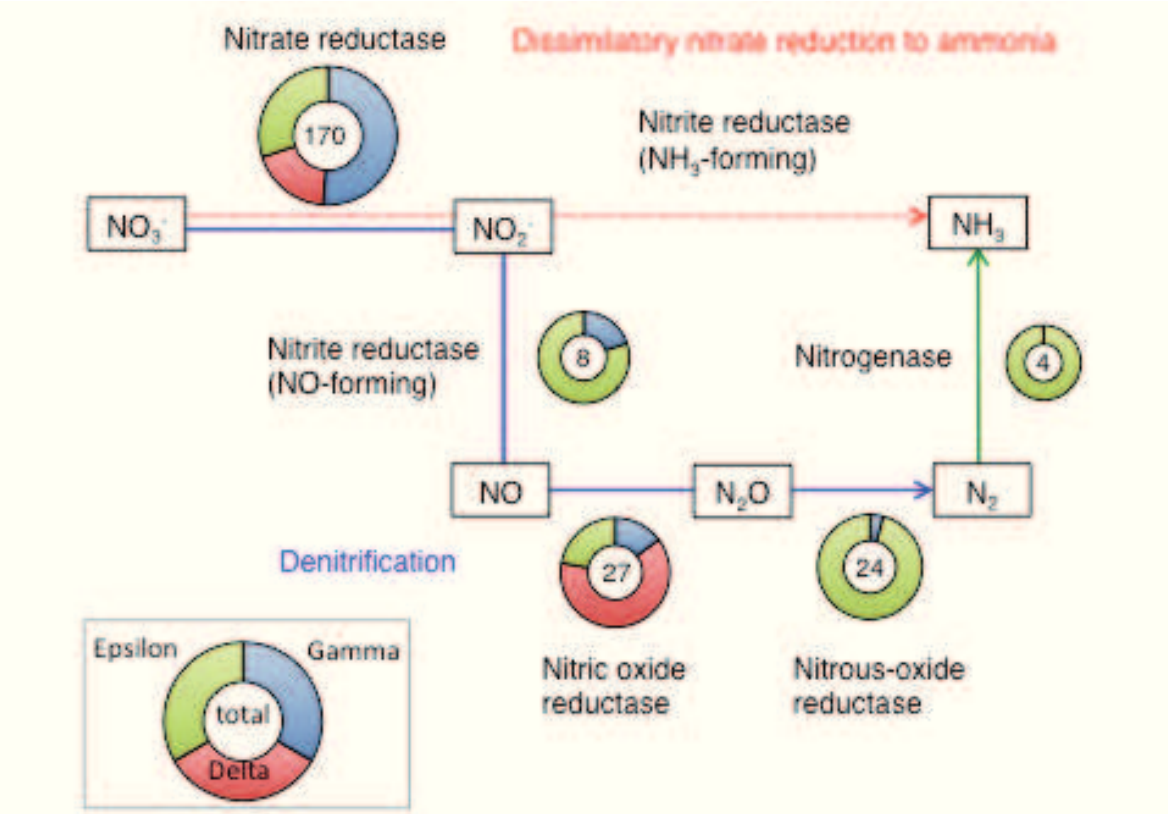


Figure 2: Components of the dissimilatory nitrogen metabolism pathway identified in this study. The doughnut-like chart indicates the proportion of each taxonomic bin. The number in the chart shows the sequence reads of the corresponding enzymes.

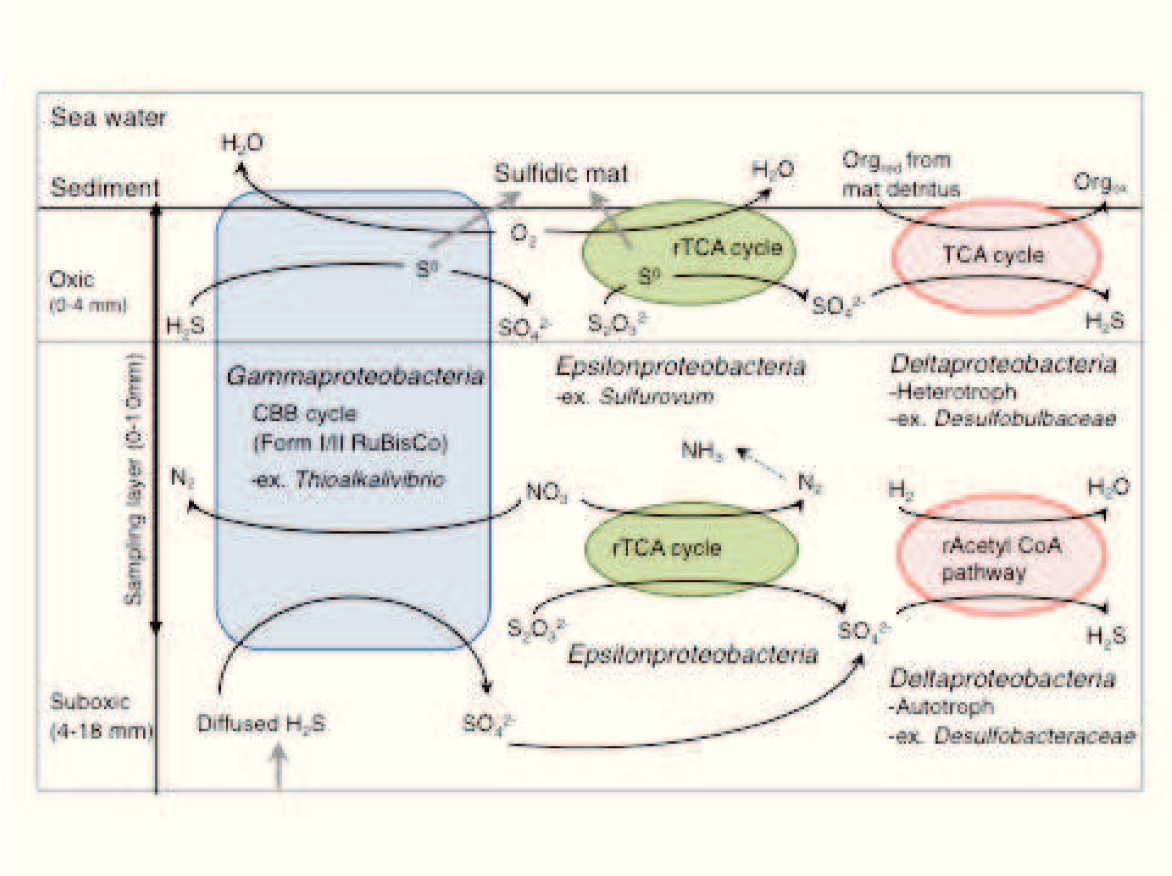


Figure 3: Model of the energy metabolism and spatial distribution of microorganisms in the sediment surface layer of site F. The reduced sulfur diffused from hydrothermal vent field is oxidized to sulfate by *G*ammaproteobacteria and *E*psilonproteobacteria . Subsequently, deltaproteobacterial autotroph reduces sulfate to sulfide at deeper anoxic sediment. *G*ammaproteobacteria and *E*psilonproteobacteria which are mainly consists of genus Sulfurovum and accumulate elemental sulfur forming sulfidic mat, also oxidize the reduced sulfur compounds at the upper sediment. Sequential sulfate reduction by heterotrophic *D*eltaproteobacteria is also occurred at the top sediment.

SUPPLEMENTARY TEXT

Metagenomic fosmid library construction

Sediment from site F (0-1 cm) at the Logatchev hydrothermal vent field was used for fosmid library construction. DNA was extracted as described previously (Zhou et al., 1996), followed by gel purification (Plaque GP Agarose, Biozym, Hess. Oldendorf, Germany) to remove humic substances. The fosmid library was constructed using the EpiFOS CopyControl Fosmid Library Production Kit (Epicentre, Madison, WI) according to the manufacturer's instructions with the modification that the DNA was concentrated using MICROCON YM-100 columns (Promega, Mannheim, Germany). Size-selection was performed by pulsed-field gel electrophoresis (PFGE) to obtain appropriate DNA fragment lengths (approximately 40 kb) for cloning.

Screening for fosmids carrying phylogenetic and metabolic key genes

2,304 randomly chosen fosmid clones were screened for genes of the 16S rRNA, the alpha subunit of the dissimilatory adenosine-5'-phosphosulfate (APS) reductase (□□□A), and the large hydrogenase subunit (hupL). Fosmids with 16S rRNA genes were identified using universal primer sets for *Bacteria* and *Archaea* as described previously (Schauer et al., 2011). The primer sets AprA-1-FW (TGGCAGATCATGATYMAYGG) and AprA-5-RV (GCGCCAACYGGRCCRTA) (Meyer and Kuever, 2007), and HUPLX1 (5'-GACCCSGTBACSCGNATYGARGG-3') and HUPLW2 (5'-RCANGCNAGRCASGGGTCGAA-3') (Csaki et al., 2001) were used for amplification of □□□A and hupL genes, respectively. The PCR reaction contained 10 - 100 ng template DNA, 0.5 µM of each primer, 10 mM dNTPs, 1 x buffer, 1 x enhancer, and 5 U of the Eppendorf-Taq DNA Polymerase (Eppendorf, Hamburg, Germany). After initial 3 min. at 94 □C each cycle consisted of 1 min. at 94 □C, 1 min at 58 □C annealing, and 3 min. elongation at 72 □C. The amplicons were purified using a PCR purification kit (QIAGEN, Hilden, Germany), and cloned using the TOPO TA Cloning Kit for sequencing (pCR4-TOPO) (Invitrogen, Karlsruhe, Germany). Clones with a correct insert size were sequenced using the vector primer M13 R. Twelve selected fosmids were Sanger sequenced on a 3130xl capillary DNA sequencer (Applied Biosystems, Carlsbad, CA) generating 12FOSMID assembly.

Screening for fosmids based on the phylogenic analysis from metagenomic library

The randomly selected fosmid clones from the library were end-sequenced with Sanger-sequencing. This yielded 6,582 bi-directional end-sequence. These were taxonomically assigned as described previously (Hanno Teeling et al., 2012). Based on

these results, 42 fosmids were selected and subjected to 454 pyrosequencing. Assembly and was conducted independently by Newbler and Mira as described in materials and methods. We also combined both assemblies to a consensus using the Minimus2 program generating 42FOSMID assembly.

G□□□ □□□□□□□□□□, □□□□□□□□□□

Metagenome sequences were analyzed with the GenDB v2.2 annotation system (Meyer □□□□, 2008). Genes were called with the gene prediction programs GLIMMER 3.02 (Delcher □□□□, 2007), MetaGene 1.0 (Noguchi □□□□, 2006), ZCURVE 1.02 (Guo □□□□, 2003), and MED 2.0 (Zhu □□□□, 2007). All predicted coding open reading frames were subjected to similarity searches against sequence databases [NCBI-nr, UniPROT/SWISSPROT (Apweiler □□ □□, 2001)], protein family databases [Pfam (Sonnhammer □□□□, 1997), InterPro (Apweiler □□□□, 2000)], and COG (Tatusov □□□□, 1997), KEGG (Kanehisa □□□□, 2002)], as well as signal peptide [SignalP v2.0, (Nielsen □□□□, 1999)] and transmembrane helix predictions [TMHMM v2.0, (Krogh □□□□, 2001)]. Subsequently, an initial automatic annotation was generated by the use of fuzzy logic-based MicHanThi (Quast, 2006). Genes of interest were manually curated using JCoast v1.6 (Richter □□□□, 2008).

□□□□□□□□ □□ □□□□□□□□□□□□□□□ □□□□□□□□□□ □□□□□□ □*HF-*□ □□□□ □□□□□□ □□□

Assembled contigs of the LHF-Meta assembly were taxonomically classified using a modified pipeline published elsewhere (Ferrer □□ □□, 2012; Teeling □□ □□, 2012). In brief, the taxonomic classification of contigs was predicted as follows: a consensus from four individual taxonomic prediction tools was used in order to infer the taxonomic affiliation of the contigs: (I) CARMA (Krause □□□□, 2008) infers taxonomy of sequences by post-processing genes with HMMER hits to the Pfam database. (II) KIRSTEN (Kinship Relationship Reestablishment unpublished) infers taxonomy of sequences by post-processing BLAST hits. (III) analysis of full and partial 16S rRNA genes and (IV) mapping of the contigs on a well-chosen set of 339 marine reference genomes taken from EnvO-lite environmental ontology (Hirschman □□ □□, 2008). The final logic consolidates the individual tool□s taxonomic predictions into a consensus using a weighted assessment on all existing 27 ranks of the NCBI taxonomy from superkingdom to species.

KEGG orthology analysis with an expectation value cut-off of < E-15 indicated that the majority of genes with KEGG assignment stemmed from *G*□□ □ □□□□□□□□□□□□□□□□□□ (33.4% of total reads mapping on contigs with ORFs with taxonomic assignment),

*D*eltaproteobacteria (15.2%), *E*psilonproteobact ria (6.4%), *F*irmicutes (7.5%) and *B*acteroidetes (4.6%) (Supplementary Table 2, Supplementary Figure 6).

*I*ssues of microbial community analyses tools

Several reasons might contribute to the partly large differences between the methods used to quantify microbial groups at Site F. Group-level CARD-FISH probes often exhibit lower group coverage and more outgroup hits in comparison to domain-specific probes (Amann and Fuchs, 2008). Probes targeting *B*acteroides (CF319a: outgroup hits: 5%) and the *D*eltaproteobacteria [outgroup hits of DELTA495a: 17%; DELTA495b: 71%; DELTA495c: 24%; (L cker et al , 2007)] have many outgroup hits, which can result in overestimating the corresponding abundances. In contrast, the *G*ammaproteobacteria probe GAM42a has a low group coverage (76%), as it does for example not cover the family Xanthomonadacea (Amann and Fuchs, 2008: Yilmaz et al , 2006), which was detected in the LHF metagenome by tag-sequencing analysis (5, 42, and 6 sequences in tag_gm1, tag_gm3 and tag NR15, data not shown). Different results of the taxonomic sequence classification and the 16S rRNA analysis also may result from biases, such as the low number of analyzable 16S rRNA reads (0.05% of all reads), which is statistically critical, especially for groups of lesser abundance. Also varying numbers of rRNA operons in microbial species skew the results of 16S rRNA fragment analysis. Moreover, public sequence databases for protein-coding genes used for sequence classification are still taxonomically biased, which hampers classification of taxa with no sequenced close relative. Finally, the different sample preparations for CARD-FISH and metagenome analysis introduced biases at several steps (cell dissolution from sediment, cell permeabilization, DNA extraction, DNA sequencing), so that CARD-FISH and metagenome analyses essentially shed a different light on a sample s true′ microbial biodiversity. If all methods or at least some independent ones, point to the same direction, the chance is high, that the abundance of a microbial community has been pretty well assessed. If not, further analyses are necessary. Here, likely CARD-FISH with several probes specific for the taxonomic group that has to be investigated, may be the most promising but time consuming approach.

P*C*R amplification and phylogenetic analysis of genes for sulfur metabolism

The genes for the dissimilatory adenosine-5'-phosphosulfate (APS) reductase (*A*) and the SoxB component of the sulfur oxidation enzyme complex (sox *B*) were amplified (*A* primers: AprA-1-FW (5 -TGGCAGATCATGATYMAYGG-3) and AprA-

5-RV (5´-GCGCCAACYGGRCCRTA-3´) (Meyer and Kuever, 2007); □□□*B* primers: and soxB432F (5´-GAYGGNGGNGAYACNTGG-3´) and SoxB1446B (5´-CATGTCNCCNCCRTGYTG-3´) (Petri □□ □□., 2001). The PCR reaction contained 10 - 100 ng template DNA, 0.5 □M each primer, 10 mM dNTPs, 1 x buffer, 1 x enhancer, and 5 U of Taq DNA Polymerase (Eppendorf, Hamburg, Germany). After an initial denaturation step for 3 min. at 94 □C, each cycle consisted of 1 min. at 94 □C, 1 min at 58 □C (□□□*A*) or 55 □C (□□□*B*), and 3 min. at 72□C. The amplicons were purified using a PCR purification kit (QIAGEN, Hilden, Germany), and cloned using the TOPO TA Cloning Kit for sequencing (pCR4-TOPO) (Invitrogen, Karlsruhe, Germany). Clones with correct insert sizes were sequenced using the vector primer M13 R. The resulting sequences were translated and analyzed using the ARB software package (Ludwig □□ □□., 2004). 88736.

Refe□e□ce□

Amann, R., and Fuchs, B. M. (2008). Single-cell identification in microbial communities by improved fluorescence in situ hybridization techniques. Nature Reviews Microbiology **6**: 339-348.

Apweiler, R., Attwood, T. K., Bairoch, A., Birney, E., Biswas, M., Bucher, P., et al (2000). InterPro– n integrated documentation resource for protein families, domains and functional sites. *B*ioinformatics **16**: 1145-1150.

Apweiler, R., Attwood, T. K., Bairoch, A., Birney, E., Biswas, M., Bucher, P., et al (2001). The InterPro database, an integrated documentation resource for protein families, domains and functional sites. Nucleic acids research **29**: 37-40.

Csaki, R., Hanczar, T., Bodrossy, L., Murrell, J. C., and Kovacs, K. L. (2001). Molecular characterization of structural genes coding for a membrane bound hydrogenase in Methylococcus capsulatus (Bath). *FE*□ □ microbiology l]tters **205**: 203-207.

Delcher, A. L., Bratke, K. A., Powers, E. C., and Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. *B*ioinformatics **23**: 673-679.

Ferrer, M., Werner, J., Chernikova, T. N., Bargiela, R., Fernandez, L., La Cono, V., et al (2012). Unveiling microbial life in the new deep-sea hypersaline Lake Thetis. Part II: a metagenomic study. *E*nviron Microbiol **14**: 268-281.

Guo, F. B., Ou, H. Y., and Zhang, C. T. (2003). ZCURVE: a new system for recognizing protein-ɔding genes in bacterial and archaeal genomes. Nucleic acids research **31**: 1780-1789.

Hedderich, R., Klimmek, O., Kr□er, A., Dirmeier, R., Keller, M., and Stetter, K. O. (2006). Anaerobic respiration with elemental sulfur and with disulfides. *FE*□ □ microbiology reviews **22**: 353-381.

Hirschman, L., Clark, C., Cohen, K. B., Mardis, S., Luciano, J., Kottmann, R., et □□. (2008). Habitat-Lite: a GSC case study based on free text terms for environmental metadata. OM *IC*S *A J*ournal of *I*ntegrative *B*iology **12**: 129-136.

Kanehisa, M., Goto, S., Kawashima, S., and Nakaya, A. (2002). The KEGG databases at GenomeNet. Nucleic acids research **30**: 42-46.

Krause, L., Diaz, N. N., Goesmann, A., Kelley, S., Nattkemper, T. W., Rohwer, F., et al (2008). Phylogenetic classification of short environmental DNA fragments. Nucleic acids research **36**: 2230-2239.

Krogh, A., Larsson, B., Von Heijne, G., and Sonnhammer, E. L. L. (2001). Predicting transmembrane protein topology with a hidden markov model: application to complete genomes1. *J*ournal of molecular biology **305**: 567-580.

L□ker, S., Steger, D., Kjeldsen, K. U., MacGregor, B. J., Wagner, M., and Loy, A. (2007). Improved 16S rRNA-targeted probe set for analysis of sulfate-reducing bacteria by fluorescence in situ hybridization. *J*ournal of microbiological methods **69**: 523-528.

Ludwig, W., Strunk, O., Westram, R., Richter, L., Meier, H., Buchner, A., et al (2004). ARB: a software environment for sequence data. Nucleic acids research **32**: 1363-1371.

Meyer, B., and Kuever, J. (2007). Molecular analysis of the distribution and phylogeny of dissimilatory adenosine-5'-phosphosulfate reductase-encoding genes (aprBA) among sulfur-oxidizing prokaryotes. Microbiology **153**: 3478-3498.

Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E. M., Kubal, M., et al (2008). The metagenomics RAST server- public resource for the automatic phylogenetic and functional analysis of metagenomes. *B□C* bioinformatics **9**: 386.

Nielsen, H., Brunak, S., and Von Heijne, G. (1999). Machine learning approaches for the prediction of signal peptides and other protein sorting signals. Protein *E*ng□neering **12**: 3-9.

Noguchi, H., Park, J., and Takagi, T. (2006). MetaGene: prokaryotic gene finding from environmental genome shotgun sequences. Nucleic acids research **34**: 5623-5630.

Petri, R., Podgorsek, L., and Imhoff, J. F. (2001). Phylogeny and distribution of the soxB gene among thiosulfate-oxidizing bacteria. *FE*□□ microbiology letters **197**: 171-178.

Quast, C. (2006). MicHanThi-design and implementation of a system for the prediction of gene functions in genome annotation projects. Universitä *B*remen , *D*iploma work ,.

Richter, M., Lombardot, T., Kostadinov, I., Kottmann, R., Duhaime, M. B., et al (2008). JCoast- biologist-centric software tool for data mining and comparison of prokaryotic (meta) genomes. *B□C* bioinformatics **9**: 177.

Sonnhammer, E. L. L., Eddy, S. R., and Durbin, R. (1997). Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins Structure *F*unction and *G*enetics **28**: 405-420.

Tatusov, R. L., Koonin, E. V., and Lipman, D. J. (1997). A genomic perspective on protein families. Science , *278*(5338), 631-637.

Teeling, H., Fuchs, B. M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C. M., et al (2012). Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. Science **336**: 608-611.

Yilmaz, L. S., Okten, H. E., and Noguera, D. R. (2006). Making all parts of the 16S rRNA of Escherichia coli accessible in situ to single DNA oligonucleotides. *A*ppl *E*nviron Microbiol **72**: 733-744.

Zhou, J., Bruns, M. A., and Tiedje, J. M. (1996). DNA recovery from soils of diverse composition. *A* [illegible], **62**: 316-322.

Zhu, H., Hu, G. Q., Yang, Y. F., Wang, J., and She, Z. S. (2007). MED: a new non-supervised gene prediction algorithm for bacterial and archaeal genomes. *B* [illegible] **8**: 97.

Supplementary Tables and figures

Supplementary Table 1 General features of the metagenome Logatchev site F

LHF-454 dataset (Raw reads)	
Total reads	1,152,840
Total length (Mbp)	408
Mean reads (bp)	354
Median reads (bp)	391
De-replicated dataset	
Total reads	757,646
Total length (Mbp)	277.5
16S rRNA genes	369 (0.05%)[a]
23S rRNA genes	714 (0.09%)[a]
Assembled sequence information	
Total contigs	151,746
Total length (Mbp)	83.7
Maximum contig size	27,087
N_{50} of contigs	509

[a]percentage in the de-replicated total reads

Supplementary Table 2

Number of LHF-454 sequence reads annotated in autotrophic carbon metabolism.

	Total number of sequence reads	*Gamma-proteobacteria*	*Delta-proteobacteria*	*Epsilon-proteobacteria*
Calvin-Benson-Bassham cycle				
RuBisCO (ribulose biphosphate carboxylase)	83	9	0	0
phosphoglycerate kinase	94	30	2	7
glyceraldehyde 3-phosphate dehydrogenase	163	61	21	11
triose-phosphate isomerase	68	17	1	7
fructose-bisphosphate aldolase / sedoheptulose-1,7-bisphosphate aldolase	130	53	19	4
fructose 1,6-bisphosphatase	22	3	2	0
D-Fructose 6-phosphate:D-glyceraldehyde-3-phosphate glycolaldehyde transferase	168	73	11	24
pentose-5-phosphate 3-epimerase	55	15	12	3
phosphoribulokinase	30	29	0	0
fructose-bisphosphate aldolase / sedoheptulose-1,7-bisphosphate aldolase	93	55	0	0
D-fructose 1,6-bisphosphatase class 2/sedoheptulose 1,7-bisphosphatase	22	3	2	0
glycolaldehydetransferase	168	73	11	24
ribose-5-phosphate isomerase	71	30	2	6
Reductive TCA cycle				

isocitrate dehydrogenase	62	19	7	14
aconitase	158	66	26	11
ATP-citrate lyase	16	0	0	10
pyruvate:ferredoxin oxidoreductase	155	7	13	21
phosphoenolpyruvate synthase	153	28	46	10
phosphoenolpyruvate carboxylase	9	8	0	0
malate dehydrogenase	9	12	14	3
fumarate hydratase	83	21	12	14
fumarate reductase	250	46	58	55
succinyl-CoA synthetase	109	29	10	13
2-oxoglutarate:ferredoxin oxidoreductase	134	0	37	13
□ed□c i□e ace□□□c□e□□□□ e A □a□h□a□				
formate dehydrogenase (NADP+)	310	120	77	7
formatetetrahydrofolate ligase	72	3	37	0
methenyltetrahydrofolate cyclohydrolase	43	11	9	6
methylenetetrahydrofolate dehydrogenase (NADP+)	39	7	2	6
methylenetetrahydrofolate reductase	33	18	71	1
CO-methylating acetyl-CoA synthase	224	0	19	0
3-h□d□□□□□□□i□□a□e c□c□e				
malony- CoA reductase (malonate semialdehyde-forming)	0	0	0	0
3-hydroxypropionate dehydrogenase (NADP+)	0	0	0	0
3-hydroxypropionyl-CoA synthase	0	0	0	0
3-hydroxypropionyl-CoA dehydratase	0	0	0	0
acrylyl-CoA reductase (NADPH)	0	0	0	0
propionyl-CoA carboxylase	65	0	12	0
methylmalonyl-CoA epimerase	0	0	0	0
methylmalonyl-CoA mutase	165	1	39	0
L-malate CoA transferase	0	0	0	0
malyl-CoA lyase	0	0	0	0
acetyl-CoA carboxylase	170	102	12	21

The key enzymes are shown in gray.

Supplementary Table 3

The number of LHF-454 sequence reads in each KEGG classification in taxonomically binned LHF metagenome

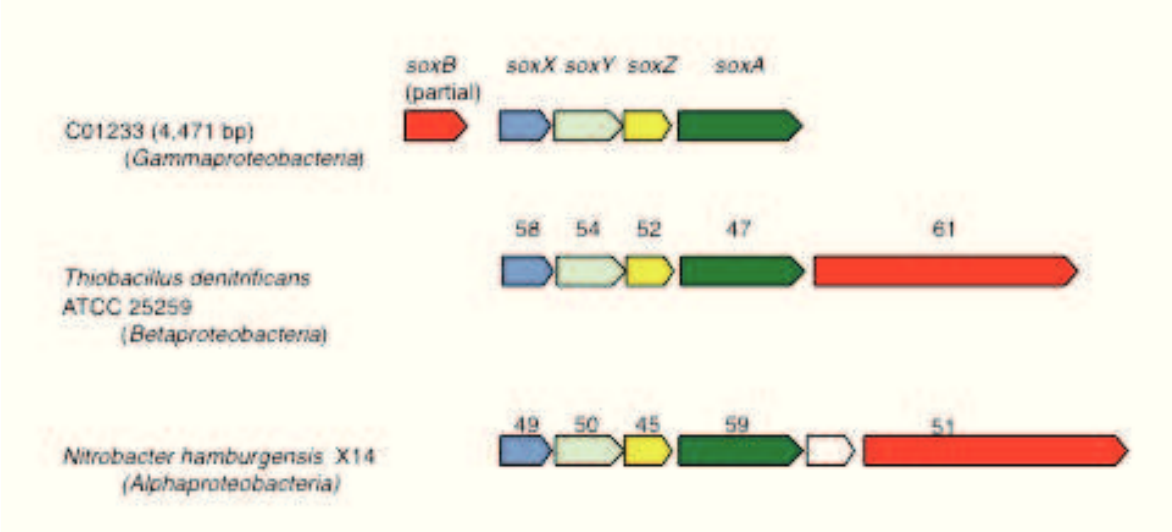
	whole taxon	*G*[illegible]	*D*[illegible]	*E*[illegible]	*F*[illegible]	*B*[illegible]
Carbohydrate metabolism	19161	5377	3144	1095	1503	1006
Starch and sucrose metabolism	1465	695	219	18	46	81
Amino acid metabolism	16566	5342	2445	1202	1204	866
Metabolism of Other Amino Acids	3552	1340	533	205	265	184
Energy Metabolism	13472	4209	2176	947	943	486
CO_2 Fixation	1588	307	301	171	135	72
Sulfur Metabolism	663	239	137	55	79	45
Nitrogen Metabolism	2297	556	212	206	138	87
Methane Metabolism	3377	765	757	164	249	72
Nucleotide Metabolism	8824	3092	1288	731	713	478
Metabolism of Cofactors and Vitamins	5579	2084	768	421	476	298
Lipid Metabolism	4050	1060	584	405	362	206
Fatty acid Metabolism	756	144	224	21	105	49
Glycan Biosynthesis and Metabolism	1867	620	301	211	149	112
Xenobiotics Biodegradation and Metabolism	2763	651	438	108	275	124
Benzoate Degradation	419	59	116	2	61	14
Biosynthesis of Other Secondary Metabolites	2173	823	316	68	98	113
Metabolism of Terpenoids and Polyketides	1819	584	238	125	116	127
Enzyme Families	1064	286	289	45	48	44
Replication and Repair	7793	3158	1136	482	857	328
Translation	5791	2243	807	405	419	263
Folding, Sorting and Degradation	3057	1275	509	278	132	92
Transcription	1253	448	190	83	69	91
Membrane Transport	5706	1856	698	218	310	82
Signal Transduction	2217	914	354	134	114	48
Cell Motility	397	100	81	37	46	3
Cell Growth and Death	209	109	23	12	15	25
Tatal (%)	111177	37176 (33.2)	16925 (15.2)	7164 (6.4)	8290 (7.5)	5149 (4.6)

Supplementary Table 4

Supplementary Table 5

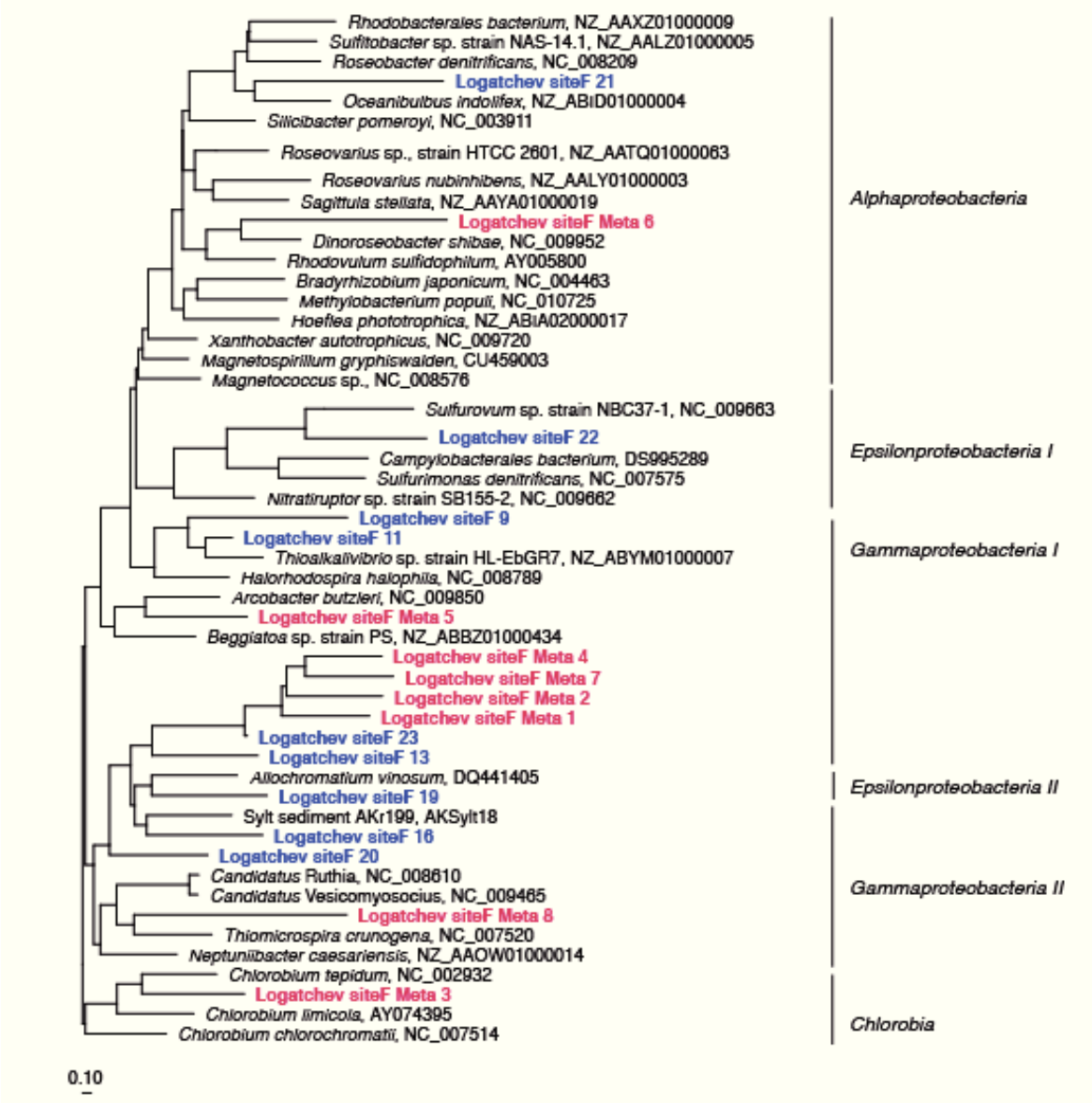
Metagenomes used for comparative analyses.

H□d□□□he□□ a□ □e□□fie□d	**L□ga□che□**	**L**ost **Ci□□**	**M□□h□a**	**Dee□ Ab**yss
Mid-Ocean Ridge	Mid Atlantic Ridge	Mid Atlantic Ridge	Juan de Fuca Ridge	-
Geological setting	Ultramafic-hosted	Ultramafic-hosted	Basalt-hosted	-
Sample	sediment (0-1cm) covered by sulfur-mat at siteF	carbonate chimney biofilm	Black smoker sulfide chimney 4143-1	Sea water from 4,000 m of Hawaii Ocean
Fluid data[a]				
Temperature (□C)	300-350	40-90	300-350	-
pH	low	9-11	2-3	-
chemical contents	hydrogen (up to 3.5 mM) methane (up to 19 mM) hydrogen sulfide (2.5 mM)	hydrogen (1-15 mM) methane (1-2 mM) hydrogen sulfide (< 2.8 mM) nearly devoid of CO2	hydrogen (0.1 mM) methane (0.05-1 mM) hydrogen sulfide (2-8 mM)	-
DNA	bulk genome	pUC18	fosmid	fosmid
Accession number		ACQI01006325-01026573	SRA009990.1	DU731018-796676
Sequencing method	454	Sanger	454	454
MG-RAST number	4496846.3	4461585.8	4497054.3	4441056.3
Raw read number	1,152,840	46,360	578,567	11,223
Total length (Mbp)	241.7	34.6	75.2	11.1
Predicted ORF number	570,224	45,725	141,258	8,289
Microbial diversity (%)[b]				
Fi □ □cutes	10.7	11	4	9
Bact roi et s	5.1	7.1	4.7	4.6
P □□t ob□ct ria	53.4	69.2	77.7	49.6
*G*ammapro eobact eria	20.7	31.2	58.7	16.7
*A*phapro e□bacteri a	6.2	19.8	8.5	21.0
*B*et pro e□bacteri a	7.8	7.4	6.4	6.4
*D*el aproteo act ri a	14.5	5.2	2.7	4.8
*E*psilonp □□t obact eria	4.0	4.9	1.0	0.5

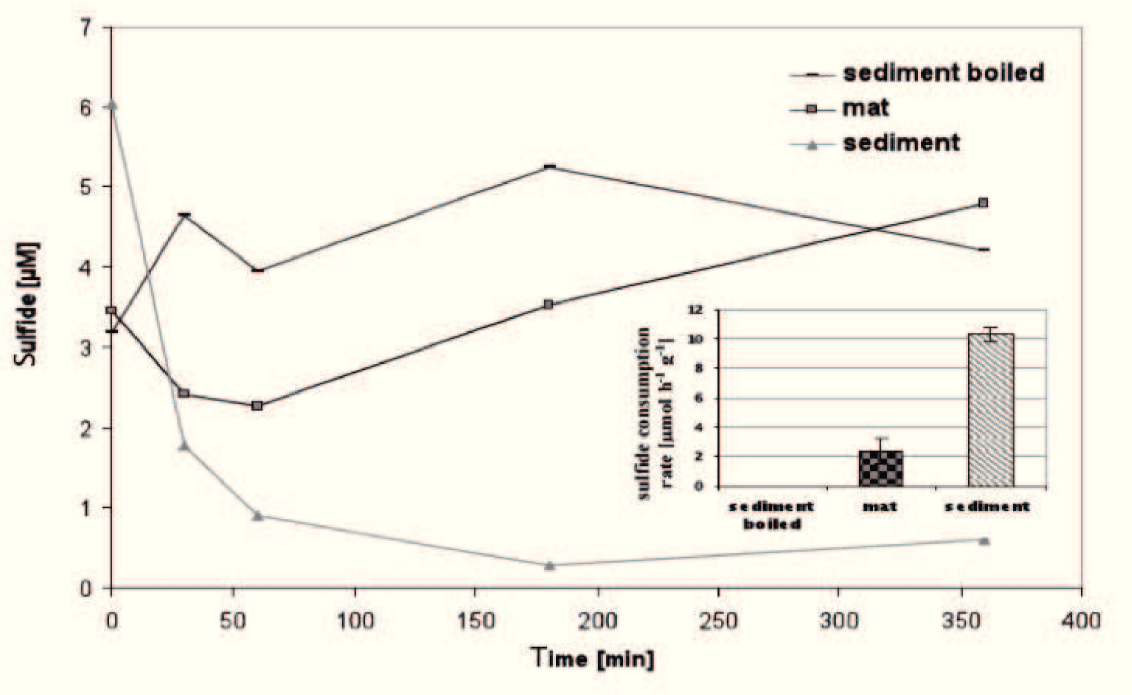


Supplementary Figure 1: Organization of genes for the Sox enzyme complex detected on a contig assigned to *G□□ □ □□□□□□□□□□□□□□□□* compared to the sox operon in the genomes of species to which the deduced amino acid sequence showed best Blast hits to. □□□*B*, sulfide thiohydrolase; □□□□, heterodimeric c-type cytochrome; □□□□, sulfur covalently-binding protein; □□□□, sulfur compound chelating protein; □□□*A* heterodimeric c-type cytochrome. Numbers given above open reading (ORF) frames of related species indicate the amino acid similarity to the respective ORF detected in the Site F metagenome.

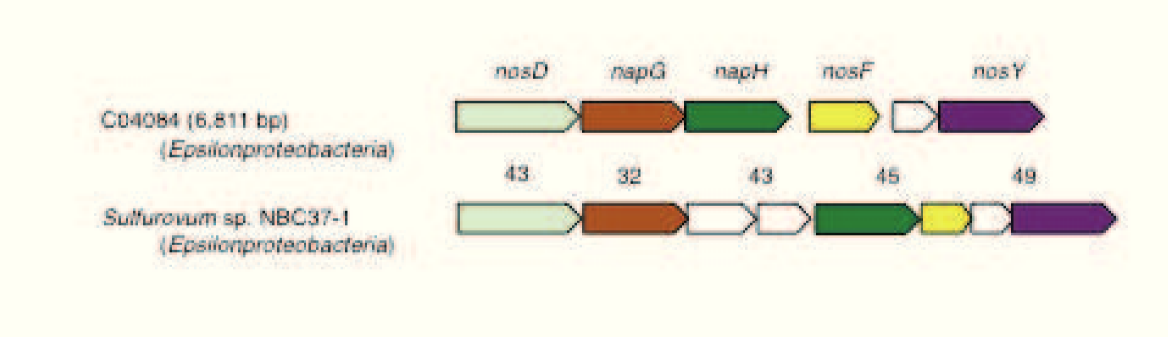
Supplementary Figure 2: Phylogenetic affiliation of AprA from Site F. Sequences deduced from the PCR and the metagenomic sequencing indicates LHF siteF (blue) and LHF siteF Meta (red), respectively. Sequences retrieved from clone libraries from the Site F are indicated in red, and sequences detected in the LHF-Meta assembly in blue. Numbers in parentheses indicate the number of identical sequences. Scale bar = 0.10 estimated substitutions per site.



Supplementary Figure 3: Phylogenetic affiliation of SoxB from Site F. Sequences deduced from the PCR and the metagenomic sequencing indicates LHF siteF (blue) and LHF siteF Meta (red), respectively. Sequences retrieved from clone libraries from the Site F are indicated in red, and sequences detected in the LHF-Meta assembly in blue. Numbers in parentheses indicate the number of identical sequences. Scale bar = 0.10 estimated substitutions per site.

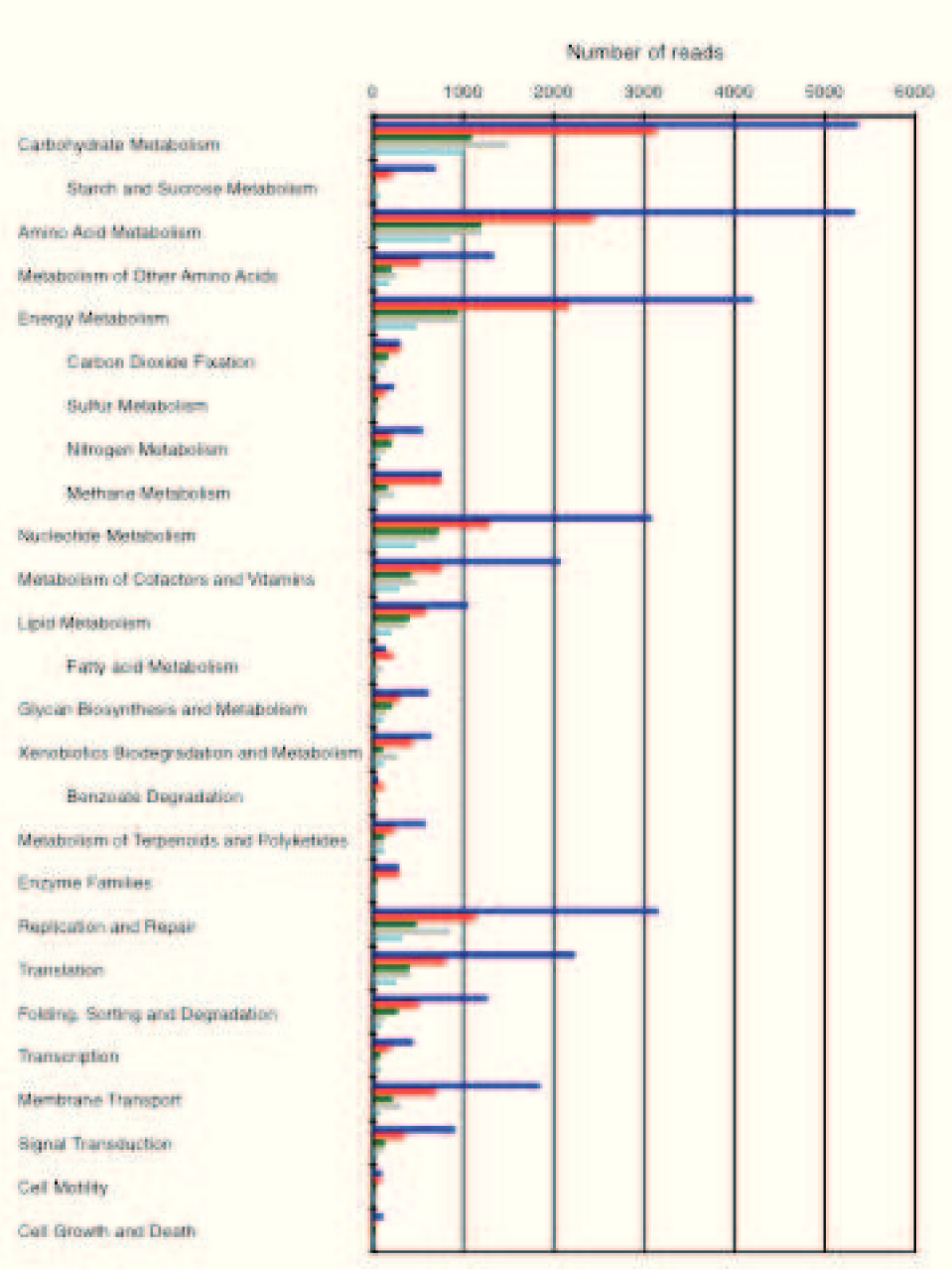


Supplementary Figure 4: Sulfide uptake and consumption rates in the surface sediment layer (0-1 cm) and in the overlying sulfur-mat at LHF.



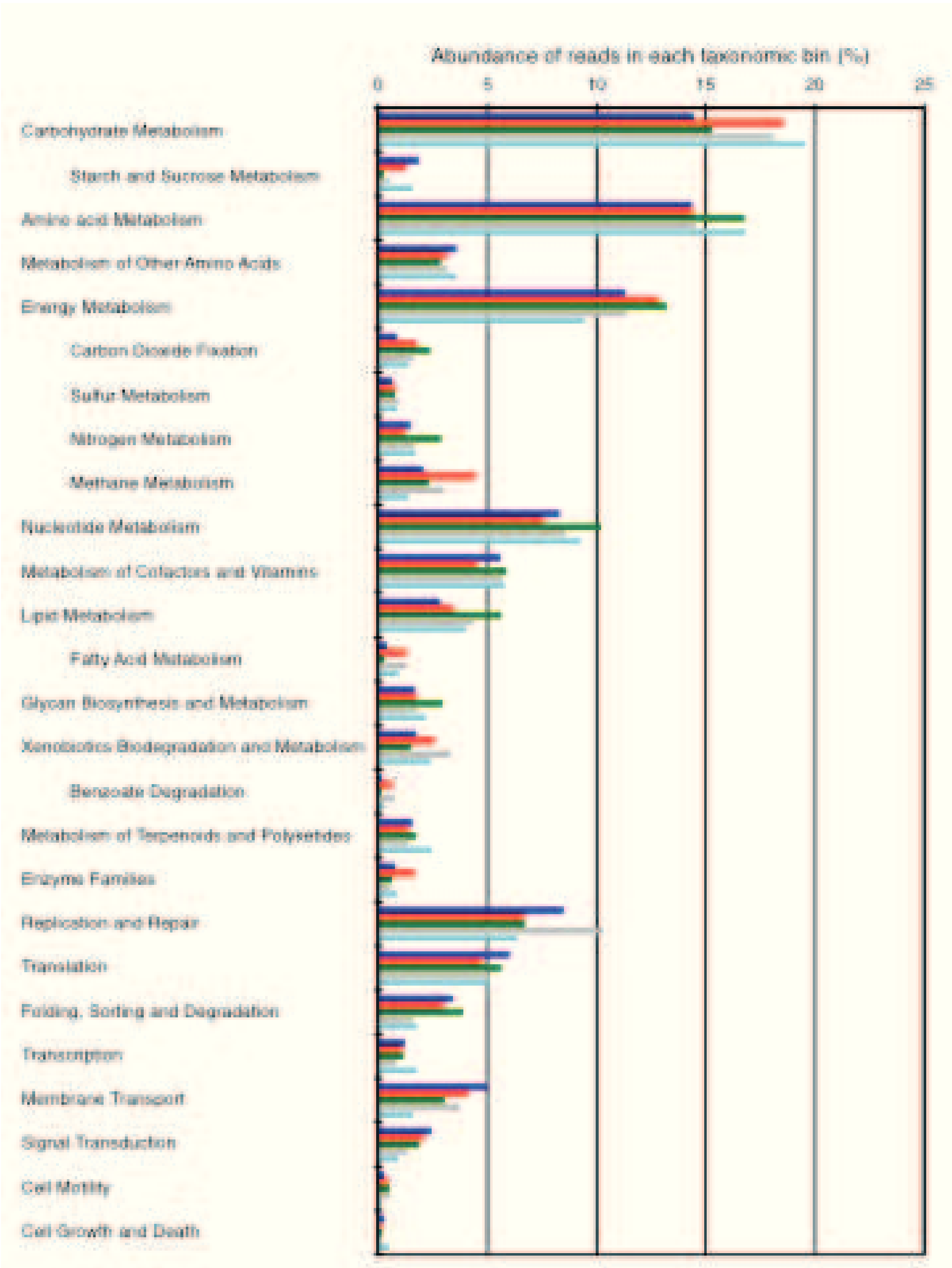
Supplementary Figure 5: Organization of genes for nitrate reduction detected on a contig assigned to *E*psilonproteobacteria compared to the sox operon in the genomes of species to which the deduced amino acid sequence showed best Blast hits to nos *DF*Y, nitrous oxide reductase maturation protein; nap *A*, nitrate reductase; nap *GH*, ferredoxin-type protein. Numbers given above open reading (ORF) frames of related species indicate the amino acid similarity to the respective ORF detected in the Site F metagenome.

Supplementary Figure 6



Distribution of abundant KEGG category classes within the taxonomically binned Site F metagenome. The number of the sequence reads in each taxonomic bin is shown. Categories in parentheses are sub-categories of higher order categories. Colors: blue: *G*ammaproteobacteria ; red*: D*eltaproteobacteria ; green: *E*psilonproteobacteria ; grey*: F*irmicutes and turquoise: *B*acteroidetes .

Supplementary Figure 7



Distribution of abundant KEGG categories classes within the taxonomically binned Site F metagenome. The proportion of the sequence reads in each taxonomic bin is shown. The sequence read counts were normalized against the total number of sequence reads within each taxobins of the LHF-454 dataset. Categories in parentheses are sub-categories of higher order categories. Colors: blue: *G*[illegible]; red: *D*[illegible]; green: *E*[illegible]; grey: *F*[illegible] and turquoise: *B*[illegible].

2.5 *H*□□□□□□□□□□ □□□□ □□□□: Complete genome sequencing and proteomics identify the first cultivated euryarchaeon from a deep-sea anoxic brine lake as polysaccharide degrader

Authors

Johannes Werner[1,2,*], Manuel Ferrer[3,*], Gurvan Michel[4], Alexander J. Mann[1,2], Sixing Huang[1], Silvia Juarez[5], Sergio Ciordia[5], Juan P. Albar[5], Mar□a Alcaide[3], Michail M. Yakimov[6], Andr□ Antunes[7], Marco Taborda[8], Milton S. da Costa[9], Rudolf I. Amann[1], Frank Oliver Gl□ckner[1,2], Olga V. Golyshina[10], Peter N. Golyshin[10,□,□], Hanno Teeling[1,□,□]: The MAMBA consortium

1 Max Planck Institute for Marine Microbiology, Celsiusstra□e 1, 28359 Bremen, Germany
2 Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen, Germany
3 CSIC, Institute of Catalysis, Marie Curie, 2, Cantoblanco, 28049 Madrid, Spain
4 UPMC University Paris 6, UMR 7139 Marine Plants and Biomolecules, Station Biologique de Roscoff, 29682 Roscoff, Bretagne, France
5 Proteomic Facility, CNB-National Centre for Biotechnology, CSIC, Darwin 3, 28049 Madrid, Spain
6 Institute for Coastal Marine Environment (IAMC), Laboratory of Marine Molecular Microbiology, CNR, Spianata S. Raineri, 86, 98122 Messina, Italy
7 Red Sea Research Center, King Abdullah University of Science and Technology (KAUST), 23955-6900 Thuwal, Kingdom of Saudi Arabia
8 Microbiology Unit, BIOCANT Biotechnological Park, 3060-197 Cantanhede, Portugal
9 Department of Life Sciences, University of Coimbra, 3001-401 Coimbra, Portugal
10 School of Biological Sciences, Bangor University, Deiniol Road, LL57 2UW Gwynedd, UK

*,□ These authors contributed equally to this work

□ Corresponding author: Peter N. Golyshin

Publication status

In preparation

My contribution

I studied the CAZyme profile of *H*. □□□□ □□□□ SARL4B^T and reconstructed the xylan □ glucose □ □-glucan pathway based on the results of Anderson □□□□ [117, 118]. I added more details to the rhodopsin proteins found in this organism. Finally, I participated in the discussion of the genomic potential of the organism.

***Environmental Microbiology* - Research Paper**

***Halorhabdus tiamatea:* Whole genome sequencing and proteomics identify the first cultivated euryarchaeon from a deep-sea anoxic brine lake as potential polysaccharide degrader**

Johannes Werner[1,2,*], Manuel Ferrer[3,*], Gurvan Michel[4], Alexander J. Mann[1,2], Sixing Huang[1], Silvia Juarez[5], Sergio Ciordia[5], Juan P. Albar[5], María Alcaide[3], Michail M. Yakimov[6], André Antunes[7], Marco Taborda[8], Milton S. da Costa[9], Tran Hai[10], Frank Oliver Glöckner[1,2], Olga V. Golyshina[10], Peter N. Golyshin[10,†,‡], Hanno Teeling[1,‡]: The MAMBA consortium

[1] Max Planck Institute for Marine Microbiology, Celsiusstraße 1, 28359 Bremen, Germany
[2] Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen, Germany
[3] CSIC, Institute of Catalysis, Marie Curie, 2, 28049 Madrid, Spain
[4] UPMC University Paris 6 and CNRS, UMR 7139 Marine Plants and Biomolecules, Station Biologique de Roscoff, 29682 Roscoff, Bretagne, France
[5] Proteomic Facility, CNB-National Centre for Biotechnology, CSIC, Darwin 3, 28049 Madrid, Spain
[6] Institute for Coastal Marine Environment (IAMC), Laboratory of Marine Molecular Microbiology, CNR, Spianata S. Raineri, 86, 98122 Messina, Italy
[7] Institute for Biotechnology and Bioengineering (IBB), Centre of Biological Engineering, Micoteca da Universidade do Minho, University of Minho, Braga, Portugal
[8] Microbiology Unit, BIOCANT Biotechnological Park, 3060-197 Cantanhede, Portugal
[9] Department of Life Sciences, University of Coimbra, 3001-401 Coimbra, Portugal
[10] School of Biological Sciences, Bangor University, Deiniol Road, LL57 2UW Gwynedd, UK
[*,‡] These authors contributed equally to this work
[†] Corresponding author: Peter N. Golyshin; phone: +44 1248 38 3629, fax: Fax +44 1248 37 0731

Running title: Ecology of *H. tiamatea* assessed by genomics and proteomics

E-mail addresses and telephone numbers of all authors:

Johannes Werner	jwerner@mpi-bremen.de	+49 421 2028 928
Manuel Ferrer	mferrer@icp.csic.es	+34 91 585 4872
Gurvan Michel	gurvan.michel@sb-roscoff.fr	+33 298 292330
Alexander J. Mann	amann@mpi-bremen.de	+49 421 2028 976
Sixing Huang	shuang@mpi-bremen.de	+49 421 2028 928
Michail M. Yakimov	michail.yakimov@iamc.cnr.it	+39 090 669 003
Silvia Juarez	sjuarez@cnb.csic.es	+34 91 585 4695
Sergio Ciordia	sciordia@cnb.csic.es	+34 91 585 4695
Juan P. Albar	jpalbar@proteored.org	+34 91 585 4696
María Alcaide	mariaalcaide@icp.csic.es	+34 91 585 4872
André Antunes	andre.antunes@ceb.uminho.pt	+351 253 601 971
Marco Taborda	taborda.marco@gmail.com	+351 231 419 040
Milton S. da Costa	milton@ci.uc.pt	+351 239 855 789
Tran Hai	t.hai@bangor.ac.uk	+44 1248 38 2566
Frank Oliver Glöckner	fgloeckn@mpi-bremen.de	+49 421 2028 970
Olga V. Golyshina	o.golyshina@bangor.ac.uk	+44 1248 38 3629
Peter N. Golyshin	p.golyshin@bangor.ac.uk	+44 1248 38 3629
Hanno Teeling	hteeling@mpi-bremen.de	+49 421 2028 976

40

Summary

Members of the euryarchaeal genus *Halorhabdus* have been found in various hypersaline habitats, but are represented by only two cultured species: *H. utahensis* AX-2^T, a facultative anaerobe from the shallow Great Salt Lake of Utah, and *H. tiamatea* SARL4B^T, an almost obligate anaerobe from the Shaban deep-sea hypersaline anoxic lake in the Red Sea. Here we present the complete genome of *H. tiamatea* SARL4B^T and elucidate its niche adaptations. *H. tiamatea* features one of the highest numbers of glycoside hydrolases in archaea and seems to have specialized in degrading recalcitrant algal and seagrass-derived hemicelluloses. Although *H. tiamatea* inhabits a temperate, dark, anoxic environment, its genome encodes genes usually found in thermophiles, light-dependent enzymes such as bacteriorhodopsin, and a complete aerobic redox chain, indicating that *H. tiamatea* evolved from an oxic habitat in the photic zone. Proteome analyses showed a pronounced stress response to oxygen, but simultaneously indicated quinone-mediated electron transport, and demonstrated the glycoside hydrolases' relevance. We could furthermore identify a closely related *Halorhabdus* member in an enrichment culture from the Mediterranean deep-sea hypersaline anoxic lake Medee. Our data support exchange between seemingly isolated hypersaline habitats and thus might explain why *Halorhabdus* species have been found in similar environments worldwide.

Introduction

Hypersaline habitats are characterized by higher than seawater salinities and are found worldwide, for example in the form of terrestrial and deep-sea brine lakes or man-made solar salterns. The salinities of these habitats range from just above

seawater to salt saturation (salinity >300), and their salt compositions range from concentrated seawater with sodium chloride as major salt (thalassohaline habitats) to compositions where other salts such as magnesium chloride dominate (athalassohaline habitats). Despite harsh conditions, microorganisms inhabit hypersaline habitats ranging from halotolerant species that merely tolerate hypersaline conditions to true halophiles that have evolutionary adapted to thrive at high salinities and require 0.5-2.5 M of salt for growth (Andrei *et al.,* 2012).

Halophiles are found in all three domains of life. In *Archaea*, halophiles are mostly present in the class *Methanococci* and the orders *Methanosarcinales* (class *Methanomicrobia*) and *Halobacteriales* (class *Halobacteria*) with its sole family *Halobacteriaceae*, which includes the majority of the known euryarchaeal halophilic species (Oren, 2008). Two major strategies have evolved to cope with high salinities and prevent enzymes from denaturing and salt-out precipitation. The first, the organic osmolyte strategy, consists of countering high osmolarities by intracytoplasmic accumulation of compatible solutes like quaternary amines (Imhoff and Rodriguez-Valera, 1984) or sugars such as trehalose (organic osmolyte strategy). The second, the salt-in strategy, relies on accumulation of high levels of internal potassium (and to lesser extents sodium) chloride.

Halophiles have been found in hypersaline habitats as different as the athalassohaline lakes of the Atacama Desert in Chile (Demergasso *et al.,* 2004), the thalassohaline Tuz Lake in Turkey (Mutlu *et al.,* 2008), the Great Salt Lake of Utah (Wainø *et al.,* 2000; Jakobsen *et al.,* 2006; Kjeldsen *et al.,* 2007), soda lakes in central Asia (Nolla-Ardèvol *et al.,* 2012), the Aran-Bidgol salt lake in Iran (Makhdoumi-Kakhki *et al.,* 2012), solar salterns (e.g. (Antón *et al.,* 1999; Antón *et al.,* 2000; Legault *et al.,* 2006; Maturrano *et al.,* 2006; Pašić *et al.,* 2007; Baati *et al.,*

2008; Tsiamis *et al.,* 2008; Boujelben *et al.,* 2012)), and even the nostril salt glands of seabirds (Brito-Echeverría *et al.,* 2009). Among the most peculiar hypersaline habitats are deep-sea hypersaline brine lakes, like the Orca Basin in the Northern Gulf of Mexico (Pilcher and Blumstein, 2007), the ice-sealed Antarctic Vida lake (Murray *et al.,* 2012), the numerous deep-sea hypersaline anoxic lakes (DHALs) in the Eastern Mediterranean Sea (Bortoluzzi *et al.,* 2011), and the Red Sea (Eder *et al.,* 1999; Eder *et al.,* 2001; Eder *et al.,* 2002).

The Shaban Deep (aka Jean Charcot Deep), a thalassohaline DHAL of the Red Sea, was discovered in 1984 (Pautot *et al.,* 1984), and since several novel species were isolated from this location, such as the gammaproteobacterial *Salinisphaera shabanensis* (Antunes *et al.,* 2003) and *Marinobacter salsuginis* (Antunes *et al.,* 2007), *Haloplasma contractile* from the novel bacterial order *Haloplasmatales* (Antunes *et al.,* 2008b; 2008) and the euryarchaeon *Halorhabdus tiamatea* (Antunes *et al.,* 2008a). *H. tiamatea* SARL4B^T stems from the brine-sediment interface of the Eastern basin of the Shaban Deep (26° 13.9' N, 35° 21.3' E, -1,447 m depth, pH 6.0, salinity: 244), and features pleomorphic (mostly rod-shaped), non-pigmented cells that grow heterotrophically under anoxic to micro-oxic conditions (optimum: 45 °C; pH 5.6 - 7.0; 27% NaCl (w/v)) (Antunes *et al.,* 2008a), but poorly under oxic conditions. *Halorhabdus*-like species have been identified by 16S rRNA gene analysis at various other sites, e.g. the Tibetian hypersaline lakes Zabuye (Fan *et al.,* 2003) and Chaka (Jiang *et al.,* 2007), the brine of the Iranian Aran-Bidgol salt sea (Makhdoumi-Kakhki *et al.,* 2012), hypersaline lakes in Mongolia (Pan *et al.,* 2006), saline hexachlorocyclohexane-contaminated soil in India (Sangwan *et al.,* 2012), the Dead Sea (Bodaker *et al.,* 2009) and Eastern Mediterranean DHALs (van der Wielen *et al.,* 2005). Apart from *H. tiamatea* SARL4B^T, there is so far only one other validly

described *Halorhabdus* species, *H. utahensis* AX-2^T (DSM 12940^T), a sediment isolate from the southern arm of the shallow thalassohaline Great Salt Lake in Utah, USA (Wainø *et al.,* 2000). *H. utahensis* AX-2^T thrives under similar conditions as *H. tiamatea* SARL4B^T (Tab. 1), also features pleomorphic (mostly rod-shaped) cells, which in contrast to *H. tiamatea* SARL4B^T are pigmented and can grow under anoxic as well as oxic conditions. Both *Halorhabdus* species exhibit a 16S rRNA sequence identity of 99.3% (Fig. 1). The genome of *H. utahensis* AX-2^T has been completely sequenced (Anderson *et al.,* 2009), whereas unitl now for *H. tiamatea* SARL4B^T only a draft sequence was available (Antunes *et al.,* 2011b). Their gene repertoires have been shown to exhibit strong overall similarities, albeit with notable differences indicating niche-differentiations, such as an increased number of genes for membrane transport and utilization of maltose, maltodextrin, phosphonate, di- and oligopeptides in *H. tiamatea* SARL4B^T (Antunes *et al.,* 2011b). *H. tiamatea* SARL4B^T and *H. utahensis* AX-2^T belong to those species within the family *Halobacteriaceae* that can degrade plant polysaccharides, like *Haloterrigena turkmenica*, and to lesser extents *Haloarcula marismortui* and *Haloferax volcanii* (Anderson *et al.,* 2011). *H. utahensis* AX-2^T has proven β-xylanase and β-xylosidase activities, which enable it to degrade abundant plant xylans (Wainø and Ingvorsen, 2003).

In this study, we sequenced and closed the genome of *H. tiamatea* SARL4B^T *de novo*, and compared it with the genome of *H. utahensis* AX-2^T with an emphasis on niche adaptation and subsequent genomic carbohydrate degradation potentials. We furthermore studied *H. tiamatea* SARL4B^T by proteomics under varying oxygen concentrations (0%, 2%, 5%). The combined results revealed insights in the lifestyle, niche differentiation and evolution of *H. tiamatea* SARL4B^T: The *H. tiamatea* SARL4B^T genome harbors 50 glycoside hydrolase (GH) genes – one of the highest

numbers reported so far in the archaeal domain. In depth analyses of these GHs indicated that *H. tiamatea* SARL4B^T has specialized on the degradation of recalcitrant algae and seagrass hemicelluloses, such as xylans and arabinans, pectins and possibly cellulose that reach the deep-sea. Further results suggest that *H. tiamatea* SARL4B^T evolved from a former thermophile that inhabited oxic waters or the oxic upper sediment of the photic zone of saline ecosystems. In addition, we identified a *Halorhabdus* species in the thalassohaline DHAL Medee (SW off the Western coast of Crete) by enrichment and subsequent draft sequencing of the 80-90% pure culture. This and other findings support transport of *Halorhabdus* spp. among seemingly isolated hypersaline habitats and thus might explain why these archaea seem to constitute a part of the autochthonous microbial community of many hypersaline habitats worldwide.

Results

Genome features of H. tiamatea SARL4B^T

The genome of *H. tiamatea* SARL4B^T (Fig. 2A-B) comprises 3,146,160 bp and consists of a chromosome (2,815,791 bp; 63.4% GC-content) and a large plasmid (330,369 bp; 57.4% GC-content). 2,744 protein-coding genes were predicted on the chromosome and 280 on the plasmid. The chromosome features a complete set of 46 tRNA synthetases for all 20 standard amino acids, one RNAseP RNA component and a single rRNA operon (Tab. 1). The *H. tiamatea* SARL4B^T chromosome exhibits a high overall level of collinearity with the *H. utahensis* AX-2^T chromosome, but differs by multiple larger inversions and minor genome rearrangements (Fig. 3A-B). Conversely, the *H. tiamatea* SARL4B^T plasmid did not show any collinearity with the *H. utahensis* AX-2^T genome.

The *H. tiamatea* SARL4B^T genome encodes at least 55 transposases on the chromosome (1.95 per 100 kbp) and 35 transposases on the plasmid (10.59 per 100 kbp). Hence the plasmid has a transposase density of 12.5%. The main chromosome features three regions with putative phage genes that are characterized by dissimilar oligonucleotide usage patterns (Fig. 2A). *H. utahensis* AX-2^T and *H. tiamatea* SARL4B^T both contain one CRISPR (clustered regularly interspaced short palindromic repeats) element. In *H. utahensis* AX-2^T, it spans 3,381 bp and includes 51 spacer elements, whereas the longer CRISPR in *H. tiamatea* SARL4B^T spans 4,703 bp with 71 spacers.

Identification of Halorhabdus sp. in the DHAL Medee

We enriched halophilic microbes from the eastern Mediterranean DHAL Medee and subsequently applied shotgun metagenomics. Taxonomic metagenome analyses revealed that the enrichment (termed ANR26) consisted of ~80-90% of *Halorhabdus* species. The longest assembled contig was 183,988 bp and featured a full-length 16S rRNA gene with 99.7% identity to *H. tiamatea* SARL4B^T and 99.4% to *H. utahensis* AX-2^T. These 16S rRNA identities exceed the 97% species criterion, above which isolates often represent a single species, but values for the Average Nucleotide Identity (ANI) between all three strains were below 87, confirming that they represent distinct species. Likewise, ANI values indicated that the ANR26 *Halorhabdus* sp. is closer related to *H. tiamatea* SARL4B^T (ANI: 86.95) than to *H. utahensis* AX-2^T (ANI: 82.95). Presence of the enriched strain in the Medee DHAL was verified by catalyzed reported deposition fluorescence *in situ* hybridization (CARD-FISH) analysis of Medee brine (Fig. 4) with newly developed *Halorhabdus*-specific probes. This analysis proved that the *in situ* abundance in the sample was

very low.

Physiology of Halorhabdus tiamatea SARL4B^T

The core metabolism of *Halobacteriaceae* including *H. utahensis* AX-2^T has recently been investigated in a comparative genomics approach (Anderson *et al.,* 2011). Hence we highlight just the major physiological traits of *H. tiamatea* SARL4B^T:

- *Monosaccharide utilization*

H. tiamatea SARL4B^T degrades hexoses via the semi-phosphorylated Entner-Doudoroff (ED) pathway to D-glyceraldehyde-3-phosphate and pyruvate. D-glyceraldehyde-3-phosphate is further oxidized to pyruvate via the lower part of the Embden-Meyerhof-Parnas (EMP) pathway. The upper part of the EMP pathway in *H. tiamatea* SARL4B^T lacks 6-phosphofructokinase. Instead, a 1-phosphofructokinase is present whose gene is co-located with that of fructose-1,6-bisphosphate aldolase. Incompleteness or variations of the EMP and gluconeogenesis pathways are often found in *Archaea*. In case of the *H. utahensis* AX-2^T, at least gluconeogenesis has been deemed non-operational due to a lack of pyruvate phosphate dikinase for phosphoenolpyruvate biosynthesis (Anderson *et al.,* 2011). This enzyme is also lacking in *H. tiamatea* SARL4B^T. The hexose fructose is funneled into the metabolism by conversion to fructose-1,6-bisphosphate (1-phosphofructokinase), and galactose is used via the Leloir pathway (Frey, 1996).

The pentose-5-phosphate (PP) pathway in *H. tiamatea* SARL4B^T is missing the oxidative branch, but the non-oxidative branch with sugar-interconversions (transaldolase, transketolase) is present. The latter is likely used to convert pentoses to fructose-6-phosphate that is subsequently converted to glucose-6-phosphate and

then passed to the semi-phosphorylated ED pathway. Without the oxidative branch of the PP pathway, glucose cannot be converted to ribulose-5-phosphate. Hence, other processes are necessary to produce this important intermediate. The genomes of both *Halorhabdus* isolates have the genes to convert xylose to ribulose-5-phosphate (xylose isomerase, xylulose kinase, ribulose-5-phosphate-3-epimerase). Furthermore, arabinose can also be converted to ribulose-5-phosphate via intermediary xylulose-5-phosphate (L-arabinose isomerase, ribulokinase, L-ribulose-5-phosphate-4-epimerase). *H. tiamatea* SARL4B^T has a ribokinase that is absent from *H. utahensis* AX-2^T, which implies that *H. tiamatea* SARL4B^T, in contrast to *H. utahensis* AX-2^T, can also use ribose in the non-oxidative PP pathway.

- *Pyruvate oxidation and fermentations*

H. tiamatea SARL4B^T relies on a fermentative lifestyle. It has a four-subunit pyruvate:ferredoxin oxidoreductase for pyruvate oxidation, which allows the disposal of reducing equivalents by releasing hydrogen from low-potential reduced ferredoxin. *H. tiamatea* SARL4B^T has genes for the biosynthesis of a cytoplasmic heterotetrameric [Ni-Fe] hydrogenase, which agrees with the observation that *H. tiamatea* SARL4B^T produces gas from sugars (Antunes *et al.,* 2008a).

H. tiamatea SARL4B^T also produces acids when grown for example on maltose (Antunes *et al.,* 2008a). It is known that this isolate features a NAD-dependent L-lactate dehydrogenase (Antunes *et al.,* 2011b) for reduction of pyruvate to L-lactate, and it has a L-lactate permease, likely for lactate export. Besides, *H. tiamatea* SARL4B^T also features a FAD-dependent D-lactate dehydrogenase. Lactate fermentation seems to be the sole mechanism in *H. tiamatea* SARL4B^T for the recycling of reduced pyridine and flavin adenin dinucleotides. Acetate is a second

likely fermentation product since the genome encodes an AMP-forming acetyl-CoA synthetase, whose reverse reaction releases acetate from acetyl-CoA while conserving energy in the form of ATP.

H. tiamatea SARL4B^T has all genes required for anaerobic glycerol degradation: a glycerol kinase and a three subunit anaerobic glycerol-3-phosphate dehydrogenase (*glpABC*) (Rawls *et al.,* 2011). In *E. coli* the anaerobic oxidation of glycerol-3-phosphate to dihydroxyacetone phosphate by GlpABC is coupled to the reduction of fumarate to succinate (Schryvers and Weiner, 1981), but other halophilic archaea such as representatives of the genera *Haloferax* and *Haloarcula* have been shown to metabolize glycerol under anoxic conditions to D-lactate, acetate and pyruvate (Oren and Gurevich, 1994).

Growth of *H. utahensis* AX-2^T is stimulated by elemental sulfur, which is reduced to hydrogen sulfide. It has been suggested that this might not be a form of respiration, but a facilitated fermentation that serves as a hydrogen sink without producing energy (Wainø *et al.,* 2000). Conversely, sulfur reduction has not been reported for *H. tiamatea* SARL4B^T (Antunes *et al.,* 2008a), and its genome does not seem to contain any respective gene(s). However, it is known that the bidirectional tetrameric hydrogenase (I) of *Pyrococcus furiosus* can also reduce sulfur (Ma *et al.,* 1993; Ma *et al.,* 2000) and since the ability for sulfur reduction is thought to be a general feature of this type of hydrogenase, sulfur reduction in *H. tiamatea* SARL4B^T might still be possible.

- Krebs cycle

H. tiamatea SARL4B^T has a complete Krebs cycle without the Krebs-Kornberg glyoxylate shunt that it mainly uses for the generation of precursors (and reducing

equivalents) for biosyntheses. Intermediates that are removed from the cycle can be replenished by carboxylation of phosphoenolpyruvate (PEP) to oxaloacetate (PEP carboxylase). This seems to be the only present anaplerotic reaction. *H. tiamatea* SARL4B^T features a malate:quinone-oxidoreductase that funnels electrons directly in the quinone pool and a ferredoxin-dependent 2-oxoglutarate oxidoreductase. Both enzymes might facilitate to run the Krebs cycle in reverse from oxaloacetate to the precursor 2-oxoglutarate, as it is known for some methanogenic archaea (Sakai *et al.,* 2011). In terms of thermodynamics, the malate:quinone-oxidoreductase can operate in reverse, albeit it does not in *Helicobacter pylori* (Kather *et al.,* 2000). *H. tiamatea* SARL4B^T lacks a distinct fumarate reductase such as the membrane-bound type found in *E. coli* or the coenzyme M-reducing cytoplasmic type found in many methanogenic archaea; hence a partially reverse Krebs cycle would involve reverse operation of the regular succinate dehydrogenase.

- *Respiration*

H. tiamatea SARL4B^T has been described as an almost obligate anaerobe that can grow under hypoxic conditions, but grows poorly under oxic conditions (Antunes *et al.,* 2008a). Nonetheless, the *H. tiamatea* SARL4B^T genome encodes the complete genes for the archaeal membrane-bound NADH:ubiquinone oxidoreductase, as well as cytochrome bd and bc ubiquinol oxidase subunits, a complete 3-subunit copper-containing cytochrome oxidase, together with cytochrome c biogenesis and copper transport genes, and a V-type ATPase.

It is known that *H. tiamatea* SARL4B^T reduces nitrate and nitrite (Antunes *et al.,* 2008a). The analysis of the *H. tiamatea* SARL4B^T genome did not reveal any membrane-associated (Nar) or periplasmic (Nap) respiratory nitrate reductase.

H. tiamatea SARL4B^T also lacks a membrane-bound Nrf-type cytochrome c nitrite reductase as it is typically found in anaerobes employing energy-conserving dissimilatory nitrate reduction to ammonium (DNRA). However, *H. tiamatea* SARL4B^T does possess genes for a Nir-type cytoplasmic nitrite reductase, which however is non-respiratory as it typically acts only as an electron sink to cope with excess reductants under anoxic conditions. Nir activity is strictly anaerobic and if constitutively required might explain the oxygen-sensitivity of *H. tiamatea* SARL4B^T.

It is not clear, whether respiration of endogenous fumarate produced by carboxylation of PEP to oxaloacetate via a reversely operating Krebs cycle is possible (see above). However, exogenous fumarate can likely be reduced to succinate as in other halophilic archaea (Oren, 1991).

- Phosphonate utilization

H. tiamatea SARL4B^T is known to harbor genes for methylphosphonate degradation (Antunes *et al.,* 2011b). The complete genome revealed all genes for methylphosphonate degradation that with the exception of *phnP* (Podzelinska *et al.,* 2009) are all arranged in a single cluster (*phnGHIJKLM*) (Kamat *et al.,* 2011). This pathway is thought to scavenge additional phosphate when inorganic phosphate is limiting and releases the methylphosphonate methyl moiety in the form of methane.

- Bacteriorhodopsin and other light-associated enzymes

Some halobacteria use light-driven ion-pumps: bacteriorhodopsin as a proton pump for chemiosmotic ATP generation and the similar halorhodopsin for importing chloride against the concentration gradient, which facilitates successive passive influx of potassium ions through a potassium channel (salt-in strategy). *H. tiamatea* SARL4B^T

has two bacteriorhodopsin genes, one bacteriorhodopsin-like gene, ten genes containing the bacterio-opsin activator domain (including one located on the plasmid). Both bacteriorhodopsins have the R-X-X-D primary proton acceptor site and D-X-X-X-K retinal Schiff-base binding site (Jiao *et al.,* 2006). One of the bacteriorhodopsins likely has a sensory function, as it is co-located with a methyl-accepting chemotaxis signal transducer gene. The second bacteriorhodopsin gene is co-located with genes encoding the bacteriorhodopsin-like protein, one of the bacterio-opsin activator domain proteins, two isoprenoid biosynthesis enzymes (likely for the retinal cofactor), and the exinuclease subunit UvrA. It is noteworthy that the genome harbors not only the complete *uvrABC* DNA repair system, but also the blue-light dependent deoxyribodipyrimidine photolyase (Park *et al.,* 1995), both of which typically restore UV-light induced DNA damages.

- Storage compounds

H. tiamatea SARL4B^T produces poly-β-hydroxy-alkanoates (PHA) as storage compounds (Antunes *et al.,* 2008a). Consequently it has genes for the reversible condensation of acetyl-CoA to PHA. A class III PHS synthase is encoded by clustered *phaE* and *phaC* genes and likely forms short chain length polymers with up to five carbon moieties in the hydroxyacyl backbone.

The *H. tiamatea* SARL4B^T genome also codes for a family 35 glycosyltransferase (GT) glycogen phosphorylase and a GH77 α-1,4-glucanotransferase which are essential for the usage of endogenous glycogen. However, *H. tiamatea* SARL4B^T lacks the usual enzymes for *de novo* glycogen biosynthesis, i.e. a glycogen synthase (GT3 or GT5) or glycogen branching enzyme (GH13) (Ball and Morell, 2003). This unexpected finding suggests that *H. tiamatea* SARL4B^T evolved a novel pathway to

generate α-1,4-glucans for carbon storage. A possible candidate gene is HTIA_0925 (Fig. S1), which is highly similar to the GH13 amylosucrase from *Neisseria polysaccharea*. The latter uses the disaccharide sucrose as sugar donor for the extension of maltodextrin, i.e. for polymerizing an α-1,4-glucan (Okada and Hehre, 1974; De Montalk *et al.,* 1999). In *H. tiamatea* SARL4B^T such a storage polysaccharide would remain linear due to the lack of any obvious glycogen branching and GH13 debranching enzyme (isoamylase). The fructose molecule that is released at every extension by the amylosucrase is likely phosphorylated by the fructokinase HTIA_1268 to directly enter glycolysis. Decomposition of the α-1,4-glucans to glucose-1-phosphate could be mediated by the combined action of the GT35 glycogen phosphorylase and a GH77 α-1,4-glucanotransferase, which finally could be converted to glucose-6-phosphate (phosphoglucomutase) for subsequent oxidation. Finally, *H. tiamatea* SARL4B^T possesses a trehalose synthase (HTIA_1280), which can convert maltose into trehalose. This non-reducing disaccharide is known to function as a compatible solute for osmosregulation, but also might serve as an additional storage compound.

- *Polysaccharide utilization*

H. tiamatea SARL4B^T and *H. utahensis* AX-2^T both contain high numbers of carbohydrate-active enzymes (CAZymes), i.e. enzymes that synthesize, modify or breakup glycosides (Henrissat and Coutinho, 2001). *H. tiamatea* SARLB^T has in total 50 GHs (15.9 GHs/Mb), 42 GHs on its chromosome and eight GHs on its plasmid (Tab. S1). This constitutes one of the highest numbers so far observed in *Archaea* (Fig. S2). According to the CAZy database (Cantarel *et al.,* 2009) as of 2012/11/10, *H. utahensis* AX-2^T has 44 GHs (14.1 GHs/Mb).

H. tiamatea SARL4B^T has genes for the degradation of xylan, arabinan, arabinoxylan and galactan-containing hemicelluloses, as well as pectin and possibly cellulose. All of these polysaccharides occur in land plant cell walls, algae (Popper *et al.,* 2011) and in seagrasses (e.g. pectin in *Zostera marina* (Zaporozhets, 2003; Khotimchenko *et al.,* 2012)). *H. tiamatea* SARL4B^T furthermore has the genomic potential to degrade exogenous storage carbohydrates such as sucrose or α-1,4-glucans (e.g. starch (Antunes *et al.,* 2008a), or glycogen).

Xylans can be hydrolyzed to xylose monomers by concerted action of seven GH10 endo-β-1,4-xylanases and three GH43 β-xylosidases. A dedicated xylose transporter can subsequently import the monomers. Arabinans can be cleaved to L-arabinose monomers by concerted action of a GH43 endo-α-1,5-L-arabinosidase and six GH51 exo-acting α-L-arabinofuranosidases. The latter can also remove decorating L-arabinose side chains from arabinoxylans and arabinogalactans. Arabinoxylans are likely degraded by concerted action of GH10 xylanases and GH51 arabinosidases. The resulting L-arabinose monomers are then likely taken up, isomerized to L-ribulose (L-arabinose isomerase) and subsequently funneled into the PP pathway. *H. tiamatea* SARL4B^T lacks any obvious galactanase and thus probably cannot use the backbones of galactans and arabinogalactans. However, its genome codes for a GH4 α-galactosidase (HTIA_1076) and a GH42 β-galactosidase (HTIA_p2897), which likely enable *H. tiamatea* SARL4B^T to cleave galactose side chains from various hemicelluloses (Popper *et al.,* 2011). These galactose monomers are subsequently metabolized via the Leloir pathway.

The *H. tiamatea* SARL4B^T genome encodes a single PL1 family polysaccharide lyase with pectate lyase function (HTIA_2451), and a GH88 enzyme (HTIA_0144). The latter is similar to the d-4,5-unsaturated β-glucuronyl hydrolase from *Bacillus* sp.

GL1, which is involved in the hydrolysis of unsaturated glycosaminoglycan (GAG) oligosaccharides released by GAG lyases (Itoh *et al.,* 2004). Similarly, HTIA_0144 likely cleaves the unsaturated oligopectins released by its sole PL.

H. tiamatea SARL4B^T comprises a modular family GH9 with a C-terminal CBM3. This architecture is reminiscent of the endo-processive cellulase E4 from *Thermomonospora fusca* (Sakon *et al.,* 1997). Therefore this enzyme may be pivotal for the degradation of crystalline cellulose. Two GH5 family glycoside hydrolases with possible glucanase functions could also have a complementary cellulolytic activity. The released oligoglucans may be further degraded by GH3 β-glucosidases (six genes). Alternatively, resulting cellobiose dimers might be processed by two GH94 cellobiose phosphorylases, which use inorganic phosphate to cleave cellobiose to glucose and glucose-1-phosphate (Yernool *et al.,* 2000). A third GH94 gene (HTIA_1257) has a sequence highly similar to the laminaribiose phosphorylase from *Paenibacillus* sp. YM1 (Kitaoka *et al.,* 2012). While *H. tiamatea* SARL4B^T lacks any obvious laminarinase, it still may use exogenous laminaribioses released by laminarin-degrading microorganisms.

H. tiamatea SARL4B^T has a putative maltose transporter and has been shown to grow on maltose (Antunes *et al.,* 2008a). The disaccharide maltose results from degradation of exogenous starch or glycogen by action of maltogenic GH13 glycoside hydrolases (Fig. S1), and can be subsequently hydrolyzed into two α-D-glucose units for oxidation in the semi-phosphorylated ED pathway. Sucrose is also a disaccharide that *H. tiamatea* SARL4B^T can potentially use due to presence of a GH32 β-fructosidase (HTIA_1411), which cleaves sucrose into glucose and fructose monomers.

In comparison, both *Halorhabdus* species are particularly rich in GH10 *xylanases*

and GH43 β-xylosidases (*H. utahensis* AX-2^T: 4x GH10 and 4x GH43; *H. tiamatea* SAR4LB^T: 7x GH10 and 3x GH43). Other polysaccharide-degrading enzymes are abundant in both halorhabdi as well (Tab. 2), for instance GH2 (e.g. β-mannosidase, β-glucuronidase) and GH3 (β-glucosidase). Likewise, both *Halorhabdus* genomes feature a GH9 enzyme that likely constitutes a modular cellulase. However, the CAZyme repertoires of both species also exhibit notable differences, as for example GH32 (β-fructofuranosidase) could only be found in *H. tiamatea* SARL4B^T. Differences were also found in the families GH5 and GH13: seven GH5 genes were identified in *H. utahensis* AX-2^T, one with proven cellulose activity (Anderson *et al.,* 2009), whereas only two GH5 genes were found in *H. tiamatea* SARL4B^T. Conversely, GH13 genes are more frequent in *H. tiamatea* SARL4B^T than in *H. utahensis* AX-2^T (7x GH13 in *H. tiamatea* SAR4LB^T vs. 1x GH13 in *H. utahensis* AX-2^T), whereas *H. utahensis* AX-2^T has two pectate lyase genes instead of only one.

One peculiarity of *H. tiamatea* SARL4B^T genome is that it has a cluster of GHs on its plasmid. This 330 kb plasmid harbors 280 genes that are mostly hypothetical and conserved hypothetical genes, 35 transposases as well as DNA-associated proteins and restriction enzymes. The few genes with metabolic functions are mostly concentrated in a single arabinan-degradation cluster of four GH51 exo-acting α-N-arabinofuranosidases and an L-arabinose isomerase, whereas the complementing GH43 endo-α-1,5-L-arabinosidase is encoded by the chromosome. In contrast, no plasmid has been described for *H. utahensis* AX-2^T.

The CAZyme profile of the DHAL Medee ANR26 enrichment (Tab. 2) correlates much better to the profile of *H. tiamatea* SARL4B^T than to that of *H. utahensis* AX-2^T, as for example the number for GH5 is low (two as in *H. tiamatea* SARL4B^T, instead

of seven as in *H. utahensis* AX-2^T) and the number of GH13 is high. On the other hand, the profiles of the DHAL Medee enrichment and *H. tiamatea* SARL4B^T are not identical, as for example the Medee enrichment lacks GH51 genes that in *H. tiamatea* SARL4B^T except one are encoded by the plasmid. This indicates that the enriched *Halorhabdus* species lack this plasmid.

Response to oxygen

Proteomics on *H. tiamatea* SARL4B^T cultures grown under increasing oxygen concentrations (0%, 2%, 5%) revealed a notable stress response. Of the 699 proteins that were identified (~24% of the cytosolic proteome), 455 were quantified using a label-dependent quantitation method (Tab. S2). Ten (2.2%) of these were induced under anoxic conditions ($\log_2$ ['oxic'/'anoxic'] ≤ 0.5). A moderate oxygen concentration of 2% led to a mild response with induction of 42 (9.2%) proteins ($\log_2$ [oxic/anoxic] ≥ 1.3), whereas the higher oxygen concentration of 5% caused a pronounced shift in the overall protein expression pattern with upregulation of 127 (27.9%) and down regulation of 101 (22.2%) proteins. A superoxide dismutase and an alkyl hydroperoxide reductase were notably upregulated. Also upregulated were a chlorite dismutase that acts against oxidative hypochlorite, two thiosulfate-sulfurtransferase-like rhodaneses that likely play a role in oxidative thiyl-radical scavenging (Remelli *et al.,* 2012), and a thioredoxin together with a thioredoxin reductase, which also can act as antioxidants. Furthermore, chaperonins such as archaeal thermosomes and a ferrochelatase were upregulated. Conversely, the energy metabolism was down regulated, most notably pyruvate:ferredoxin oxidoreductase, which seemed to act as major regulation unit in controlling the intracellular carbon flux, but also enyzmes from the semi-phosphorylated ED and

lower EMP pathways, hydrogenase components and a maltose transporter subunit. Likewise, most subunits of the NADH-quinone dehydrogenase were down regulated, with the notable exception of the subunit CD, which was upregulated. Interestingly, also the malate:quinone oxidoreductase was upregulated at 5% oxygen.

Almost half (24/50) of the glycoside hydrolases (3x GH2, 4x GH3, 1x GH4, 4x GH10, 6x GH13, 1x GH31, 2x GH43, 1x GH51, 1x GH67, 1x GH77) as well as three polysaccharide deacetylases (carboxyl esterase family 4) were unambiguously identified in proteome experiments, which stresses the relevance of CAZymes for the metabolism of *H. tiamatea* SARL4B^T. Based on a threshold of ±1.5 $\log_2$ [oxic/anoxic], 2% oxygen led to a down regulation of two GH10 glycoside hydrolases, whereas the single GH4 glycoside hydrolase was upregulated. Again, the response was much more pronounced at 5% oxygen, where seven GHs were down regulated (GH2, GH3, GH10, GH13), while a single GH3 and a deacetylase were upregulated. Thus, a major proportion of the identified glycoside hydrolases exhibited differential expression (mostly down regulation) as reaction to oxidative stress.

Discussion

The two so far known representatives of the genus *Halorhabdus* share strong similarities in their genome organizations and core physiologies, which is remarkable considering the large distance between their isolation sources (Great Salt Lake of Utah vs. the Red Sea). This indicates either that these geographically isolated locations still allow (or have allowed) for exchange (‘everything is everywhere: but the environment selects’, according to Lourens Baas Becking (O’Malley, 2008)), or that the respective *Halorhabdus* spp. have evolved slowly since they separated from their common ancestor, i.e. are subjected to high evolutionary pressure to maintain their

genetic content and organization. The latter, however, conflicts with the results of comparative genomics that show that both *Halorhabdus* species, while having retained an overall collinear genome since they deviated from their common ancestor, differ by multiple large-scale genomic inversions. Likewise, the many niche adaptations of *H. tiamatea* SARL4B^T vs. *H. utahensis* AX-2^T point towards rather active evolution. In this context it is noteworthy that the *Halorhabdus* species that dominated the ANR26 enrichment exhibited a closer relationship to *H. tiamatea* SARL4B^T than to *H. utahensis* AX-2^T, which reflects the closer geographical proximity of the former two as well as their more similar anoxic deep-sea brine habitats. Transport between these two habitats might date back to the Mio-Pliocene, when the Mediterranean and the Red Sea might have been part of the same hydrological system - a period during which the evaporites were formed that later became deep-sea hypersaline anoxic lakes (Hsü *et al.,* 1978).

The *H. tiamatea* SARL4B^T genome provides some interesting insights into the evolution of the genus *Halorhabdus* in general and *H. tiamatea* SARL4B^T in particular. First of all, as many other halophilic archaea *H. tiamatea* SARL4B^T and *H. utahensis* AX-2^T both share traits that are often associated with a thermophilic lifestyle. Both feature a four-subunit pyruvate:ferredoxin oxidoreductase, which is usually found in hyperthermophilic anaerobes such as *Pyrococcus furiosus*, *Thermococcus litoralis*, *Archaeoglobus fulgidus* and *Thermotoga maritima* (Mai and Adams, 1996) and in methanogenic archaea, many of which are also (moderately) thermophilic. In addition, *H. utahensis* AX-2^T is known to feature β-xylanases and a β-xylosidase with remarkably high temperature optima of 55/70 °C and 65 °C, respectively (Wainø and Ingvorsen, 2003). Both species also feature surprisingly high optimum growth temperatures considering their habitats (*H. utahensis* AX-2^T:

50 °C (Wainø *et al.,* 2000); *H. tiamatea* SARL4B^T: 45 °C (Antunes *et al.,* 2008a)). The presence of the *lysW* lysine biosynthesis gene in both points into the same direction: While most bacteria synthesize lysine from diaminopimelate, hyperthermophiles such as *Thermus thermophilus* and hyperthermophilic *Archaea* instead synthesize lysine from α-aminoadipate, whereby LysW stabilizes otherwise unstable intermediates (Horie *et al.,* 2009).

Secondly, while most *Halobacteria* feature an aerobic lifestyle, *H. tiamatea* SARL4B^T inhabits an anoxic habitat. Nevertheless, its genome features a complete set of genes for aerobic respiration as well as catalase and superoxide-dismutase genes. This indicates that *H. tiamatea* SARL4B^T evolved from a former aerobic species similar to *H. utahensis* AX-2^T. Growth of *H. tiamatea* SARL4B^T under oxic conditions is poor (Antunes *et al.,* 2008a) and our proteome analysis showed a pronounced stress response to 5% oxygen involving upregulation of stress proteins and down regulation of key energy metabolism proteins including the pyruvate:ferredoxin oxidoreductase, hydrogenase components and many of the glycoside hydrolases. It is noteworthy that subunits of the NADH-ubiquinone oxidoreductase subunits were mostly down regulated, which first of all indicates that electrons were funneled into the quinone pool under anoxic conditions. This is peculiar, since the fate of these electrons is unclear. While it might be possible that *H. tiamatea* SARL4B^T dispenses these electrons into electron sinking reactions, it is also possible that electrons are ultimately channeled to the cytochrome c oxidase, which – while present in the genome – was not detected in the proteome, possibly because it is an integral membrane protein. Such an oxygen reduction could serve as a means to detoxify low amounts of oxygen, however in this case, this does not work for higher oxygen concentrations where the NADH-ubiquinone oxidoreductase was

notably down regulated. Another peculiarity is that the malate:quinone oxidoreductase was up regulated at 5% oxygen. As mentioned, this membrane-bound enzyme also funnels electrons into the quinone pool, and upregulation thus seems to counteract down regulation of the NADH-ubiquinone oxidoreductase. A reasonable explanation would be that the malate:quinone oxidoreductase in *H. tiamatea* SARL4B^T can operate bi-directionally, and under oxic conditions is used reversibly for elevated endogenous fumarate respiration. The latter would scavenge electrons from the quinone pool and thus uncouple other possibly harmful electron transport processes at high oxygen concentrations.

A third peculiar result is that *H. tiamatea* SARL4B^T features light-driven bacteriorhodopsins as well as a blue-light dependent deoxyribodipyrimidine photolyase, yet lives in a light-less deep-sea environment. This indicates that *H. tiamatea* SARL4B^T evolved from a species from the photic zone, akin to its close relative *H. utahensis* AX-2^T. It might be that the respiration and light-dependent genes still have functions in *H. tiamatea* SARL4B^T, or that these genes are mostly obsolete, partly already non-functional due to mutations, and in the process of vanishing completely in the course of further evolutionary habitat-specific adaptations. The same reasoning has been brought forward with respect to the lack of pigmentation of *H. tiamatea* SARL4B^T in comparison to *H. utahensis* AX-2^T (Antunes *et al.,* 2008a).

Such habitat-specific niche adaptations between *H. tiamatea* SARL4B^T and *H. utahensis* AX-2^T are also reflected in their heavy metal exporters. For example, only *H. tiamatea* SARL4B^T features mercuric ion and arsenate reductases, which mirrors the specific ion compositions in the anoxic brine pools in the Red Sea (Antunes *et al.,* 2011a). Habitat-specific adaptations are also evident from differences

in sugar-transporters, as *H. utahensis* AX-2^T has been reported not to contain any known sugar transporter (Anderson *et al.,* 2011), whereas *H. tiamatea* SARL4B^T has transporters for ribose, xylulose as well as maltose and maltodextrin. In contrast to *H. utahensis* AX-2^T, *H. tiamatea* SARL4B^T contains a gene cluster for the degradation of methylphosphonate. This reflects that the Shaban Deep is a phosphate-limited environment (Antunes *et al.,* 2011b), which seems to have resulted in a positive selection for alternative phosphorus acquisition strategies.

One particularly interesting difference is both species' distinct carbohydrate utilization profiles that seem to reflect dissimilar availabilities of polysaccharide substrates between the Great Salt Lake and the Shaban Deep. For instance, *H. utahensis* AX-2^T has more cellulases than *H. tiamatea* SARL4B^T. On the other hand, *H. tiamatea* SARL4B^T seems to need the capacity to target a broader spectrum of polysaccharides, including xylans, arabinans, arabinoxylans, pectins, α-1,4/1,6-glucans and possibly cellulose. Another interesting aspect is that all genes but one for arabinan degradation are located on the plasmid, which indicates that this particular capacity might have been laterally acquired. This is corroborated by the fact that the GH cluster on the plasmid as well as one large GH cluster on the chromosome have a dissimilar tetranucleotide usage pattern that stand out even above the dissimilar patterns of the three putative phage-infected regions (Fig. 2A-B). Extensive acquisition of genes via lateral gene transfer is common in *Halobacteria* and likely has played a key role in their evolution from anaerobic methanogens (Nelson-Sathi *et al.,* 2012).

One of the plasmid's GH51 α-L-arabinofuranosidases was clearly expressed under oxic as well as anoxic conditions, which conflicts with previous findings that *H. tiamatea* SARL4B^T does not grow on arabinose (Antunes *et al.,* 2008a). However,

such discrepancies between genomic potential and growth experiments are common. For *Flavobacteria* it has been shown that sugar oligomers rather than monomers induce genes for polysaccharide degradation (Martens *et al.,* 2011). Hence growth experiments with monomers, while standard, might constitute an artificial situation for polymer-degraders under which they do not exhibit their normal *in situ* physiology. Habitat-specific adaptations with respect to polysaccharides are also evident from CAZyme profile analyses that show a distinct pattern for *H. utahensis* AX-2^T, whereas those of the deep-sea brine halorhabdi are notably similar. One key difference is the higher prevalence of GH13 family glycoside hydrolases in the deep-sea strains (Tab. 2). In *H. tiamatea* SARL4B^T there are seven GH13 family glycoside hydrolases, which might act in at least four different ways (Fig. S1): (i) degradation of exogenous α-glucans such as plant and algal starch, as well as exogenous bacterial glycogen by synergistic actions of a genuine GH13 α-amylase and a GH13 α-1,6-glucosidases, (ii) synthesis of a storage α-1,4-glucan from sucrose (catalyzed by a GH13 amylosucrase), (iii) possibly turnover of endogenous α-1,4-glucans, and (iv) turnover of trehalose. Trehalose constitutes an important compatible solute in many bacterial and archaeal halophiles, but usually not in haloarchaea, which use the salt-in osmoregulation strategy. Based on its genome it seems that *H. tiamatea* SARL4B^T uses both strategies, which might be clarified in a future metabolome study. Presence of trehalose in *H. tiamatea* SARL4B^T has already been hypothesized before (Antunes *et al.,* 2011b) as *H. tiamatea* SARL4B^T has a trehalose synthase-like GH13 (HTIA_0926). Trehalose synthase can convert maltose to the non-reducing disaccharide trehalose. Maltose is likely acquired from extracellular α-glucan degradation, since the decomposing GH13 α-amylase and GH13 α-1,6-glucosidases produce maltodextrins of various sizes down to maltose, which can be subsequently

imported by the maltose transporter (which was shown to be expressed). Alternatively, GH13 amylosucrase could concertedly add the glucose units of maltose and sucrose to an internal storage α-1,4 glucan (maltose + sucrose + α-1,4 glucan (n) = fructose + α-1,4 glucan (n+3)). In the latter case, the α-1,4-glucan would not only act as storage compound but also to regulate the internal pool of osmoregulating trehalose. Recycling of the α-1,4 glucan could be provided by concerted action of a GT35 glycogen phosphorylase, and a GH77 α-1,4 glucanotransferase. Almost all of these enzymes are co-located in a single cluster in the *H. tiamatea* SARL4B^T genome, which strengthens the hypothesis that they act together.

The broader polysaccharide substrate-spectrum of *H. tiamatea* SARL4B^T compared to *H. utahensis* AX-2^T likely reflects that sediments of DHALs such as the Shaban Deep probably receive only little land plant, algal and seagrass carbohydrate substrates. Respective detritus sinking in from the photic zone will be largely consumed when it reaches the deep-sea, and mostly the more recalcitrant components such as xylan and arabinan hemicelluloses, cellulose and connecting pectin will prevail. Such compounds accumulate at the boundary layer of the seawater and the denser brine of deep-sea anoxic lakes, and only little ultimately reaches the sediment. Hence it is likely that *H. tiamatea* SARL4B^T will be limited in available carbon. This might also explain the low *Halorhabdus* sp. *in situ* abundances that we observed it in the Medee DHAL (Fig. 4). In this context the ability to store PHA and possibly α-1,4-glucans might be crucial for *H. tiamatea* SARL4B^T to survive. *H. utahensis* AX-2^T in contrast stems from the sediment of the shallow Great Salt Lake of Utah. This lake has influxes from the Bear, Webber and Jordan Rivers, and thus likely features a higher availability of suitable plant material and algae substrates.

Conclusions

Our genome analyses suggest that *H. tiamatea* SARL4B^T is actively evolving towards a further adaptation to its temperate, light-less deep-sea habitat, but still carries genes indicative of its formerly different lifestyle. Even though *H. tiamatea* SARL4B^T shares a substantial part of genes with its close relative *H. utahensis* AX-2^T and has a similar core metabolism, it features notable habitat-specific niche adaptations. The most prominent is a notably broader repertoire of glycoside hydrolases, part of which might have been the result of recent gene acquisitions. Active evolution is also corroborated by multiple large-scale genomic inversions that *H. tiamatea* SARL4B^T and *H. utahensis* AX-2^T acquired since branching off from their common ancestor. Consequently, both *Halorhabdus* species do not seem to be slowly evolving, which implies that if very similar strains are found in distinct parts of the world such as the Great Salt Lake in Utah and a DHAL in the Red Sea, there must be exchange. This is corroborated by findings of members of the genus *Halorhabdus* in various distinct hypersaline habitats during the last years, and might even support the view that this genus constitutes a part of the autochthonous microbial community of many hypersaline habitats worldwide.

Experimental Procedures

Sampling and complete sequencing of H. tiamatea SARL4B^T

H. tiamatea SARL4B^T was isolated from the brine sediment interface of the eastern basin of the Shaban Deep in the Red Sea (26° 13.99' N, 35° 21.39' E, -1,447 m depth), sampled during cruise 52/3 of the R/V Meteor (Antunes *et al.,* 2008a). Details on the isolation and culture conditions were described in the original species

description (Antunes *et al.,* 2008a).

DNA extraction from a pure culture was performed using the modified phenol-chloroform extraction procedure (Urakawa *et al.,* 2010), and subsequently sequenced on a 454 FLX Ti sequencer (454 Life Sciences, Branford, CT, USA) using a standard library. In total 461,818 reads were generated and assembled with Newbler v. 2.3, which generated a first draft genome of 87 contigs (3.1 Mb). These contigs could be oriented in two large scaffolds. With the help of an end-sequenced fosmid library (768 end sequences with an avg. length of 400 bp), this draft genome was subsequently closed at Fidelity Systems (Fidelity Systems, Gaithersburg, MD, USA) using a combination of Illumina (GA II; PE library, ~160 mio reads) and Sanger sequencing. The Phred/Phrap (Ewing and Green, 1998; Ewing *et al.,* 1998) and Consed (Gordon, 2003) software package was used for sequence assembly and quality assessment in the finishing process. Repeat mis-assemblies were corrected with DupFinisher (Han and Chain, 2006), and a single scaffold was generated and verified using paired end sequencing of a fosmid library (384 clones, average insert size of approx. 34 kbp; Sanger reads from both termini). Initial gap closure was conducted by editing in Consed, and additional direct genomic sequencing reactions (Malykh *et al.,* 2004) were necessary to close the last gaps. Illumina reads were used to correct potential 454 base calling base errors and increase the consensus sequence's quality. This combination provided 197x coverage of the genome and 311x coverage of the plasmid. The error rate of the completed genome sequence is less than 1:100,000 (Phred50). The genome sequence has been deposited at the European Nucleotide Archive (ENA) with the accession numbers HF571520-HF571521.

Sampling and sequencing of the Medee ANR26 enrichment

Brine of the Mediterranean DHAL Medee was sampled on 17th of December 2007 during the SAMCA MedBio 2 cruise of the R/V *Urania* (34° 19.778' N, 22° 31.341' E, -3,040 m depth). Aboard, 10 ml of the sample were immediately incubated with 25 ml of DSM medium 141 in an anaerobic atmosphere (N_2/CO_2, 80/20) and later cultivated in the laboratory at 15 °C for six months (ANR26 enrichment). Total DNA was extracted using the G'NOME DNA extraction Kit (BIO 101/Qbiogene, Morgan Irvine, CA, USA) according to the manufacturer's instructions, and subsequently sequenced on a 454 FLX Ti sequencer (454 Life Sciences). In total 677,220 reads were generated. Assembly with Newbler v. 2.3 resulted into 1,647 contigs comprising 3,918,195 bp, with a mean contig length of 2,379 bp, 67 contigs exceeding 10 kbp and a longest contig of 183,988 bp. The metagenome sequence has been deposited at the ENA with the accession number ERP002033.

Taxonomic analysis of the Medee ANR26 enrichment metagenome

A metagenome taxonomic classification pipeline described elsewhere (Ferrer *et al.,* 2012; Teeling *et al.,* 2012) was used for taxonomic analysis of the metagenome. Of all reads, 602,188 (96.0%) were classified on superkingdom level, 602,005 (96.0%) on phylum level, 572,956 (91.3%) on class level and 568,971 (90.7%) on genus level. Of the classified reads, 82.6% were assigned to the genus *Halorhabdus*. Metagenome 16S rRNA gene fragment analysis with the SILVA NGS pipeline, v. 4.1.7 (Klindworth *et al.,* 2012) confirmed the strong dominance of *Halorhabdus*-like species in the ANR26 enrichment. Of the 390 reads carrying partial 16S rRNA genes, 295 had BLAST hits and 257 (87.1%) were assigned to the *Halorhabdus* genus, while 95 (24.4%) of all 16S rRNA gene fragments had no relative in the SILVA

database (SILVA release 108) (Quast *et al.,* 2012).

Gene prediction

We used the Rapid Annotation using Subsystem Technology (RAST) server (Aziz *et al.,* 2008) for the prediction of potential protein-coding genes of the genome of *H. tiamatea* SARL4B^T. In contrast, genes in the Medee ANR26 enrichment metagenome were called with the metagenome gene finder MetaGene (Noguchi *et al.,* 2006) in combination with a subsequent calling of ORFs exceeding 150 bp in the intergenic regions in order to ensure that almost no genes were overlooked.

Annotation

The *H. tiamatea* SARL4B^T genome was annotated with the RAST server and manually curated by a combination of a modified GenDB v. 2.2.1 annotation system (Meyer *et al.,* 2003) and JCoast v. 1.6.0 (Richter *et al.,* 2008). The Medee ANR26 enrichment metagenome was also annotated with GenDB. For each predicted gene, similarity searches were performed against the NCBI non-redundant protein database (Pruitt *et al.,* 2004), SWISSPROT (The UniProt Consortium, 2010), KEGG (Kanehisa and Goto, 2000) and against the protein family databases Pfam (Finn *et al.,* 2009), InterPro (Hunter *et al.,* 2011), COG (Tatusov *et al.,* 2003) and CAZy (Cantarel *et al.,* 2009). Signal peptides were predicted with SignalP v. 3.0 (Nielsen *et al.,* 1999; Emanuelsson *et al.,* 2007) and transmembrane helices with TMHMM v. 2.0 (Krogh *et al.,* 2001). Ribosomal RNA genes were identified via BLAST searches (Altschul *et al.,* 1997) against public nucleotide databases and transfer RNA genes using tRNAScan-SE v. 1.21 (Lowe and Eddy, 1997). The CRISPRFinder web service (Grissa *et al.,* 2007) was used for the identification of CRISPRs. Glycoside

hydrolases were subjected to an in-depth phylogenetic analysis in order to uncover the substrate-specificities of those belonging to multi-functional CAZyme families.

CARD-FISH and probe design

Presence of *Halorhabdus* spp. in the DHAL Medee was analyzed via CARD-FISH (Pernthaler *et al.,* 2002, Pernthaler *et al.,* 2004; Thiele *et al.,* 2011). In brief, after lysozyme treatment for cell permeabilization, the hybridization was carried out with a final probe concentration of 28 fmol/µl for 2 h at 35 °C. Signal amplification was done using Alexa Fluor® 488 labeled tyramide (1 µg/ml final concentration) for 45 minutes at room temperature. Both hybridization and amplification were done on glass slides in humidity chambers. All filters were counterstained with DAPI (4′,6-diamidino-2-phenylindole). The applied probes were ARCH915 (Stahl and Amann, 1991), NON338 (Wallner *et al.,* 1993) and a newly developed *Halorhabdus*-specific probe Halo178 (5' CCA GCT GGC GAG TCG TAT 3').

Halo178 and two helper probes were designed using ARB (Ludwig *et al.,* 2004) and the Silva 108 database. The helper probes (Halo178-h1: 5' CGG ACA TTA GCC TCA GTT TC 3' and Halo178-h2: 5' TTT CCG ACT CGC CGC ACT 3') were used to increase the target site accessibility (Fuchs *et al.,* 2000). The probe Halo178 was used in an equimolar mixture with the helpers at a formamide concentration of 25%. The Halo178 probe was evaluated with a high *in silico* specificity for *Halorhabdus* spp., as no other taxon was found with less than 1.0 weighted mismatches.

Proteomics

- Growth conditions

H. tiamatea SARL4B^T cultures were grown as described previously (Antunes *et al.,*

2008a) in HBM liquid medium (*Halobacteria* medium; DSMZ medium 372) in 200 ml glass vials until an optical density (600 nm) of 0.1 was reached. For anoxic conditions, the incubation vials were completely filled with medium and subsequently incubated in sealed cylinders with an anoxic gas phase (80% N_2 / 20% CO_2) and anaerobic container system sachets (MBraun UNILab, Garching, Germany). For the oxic conditions, the vials were filled to one quarter with medium and incubated in sealed cylinders with a gas phase with two and five percent oxygen, respectively.

- *Protein digestion and tagging with iTRAQ-4-plex®*

Peptide solutions were obtained as described elsewhere (Yakimov *et al.,* 2011), using 50 ug of proteins from each condition. Each peptide solution was labeled for two hours at room temperature with a half unit of iTRAQ Reagent Multi-plex kit (AB SCIEX, Foster City, CA, USA), previously reconstituted with 70 µl of ethanol. In the iTRAQ labeling, tags 114, 115 and 116 were used for 0%, 2% and 5% oxygen conditions, respectively. Afterwards, labeled samples containing the same protein content were combined and labeling reaction stopped by evaporation in a Speed Vac (Eppendorf, Madrid, Spain).

- *Peptide fractionation at basic pH and mass spectrometry analysis*

The digested, labeled and pooled samples were studied in detail by RP-LC-MALDI TOF/TOF MS as described elsewhere (Yakimov *et al.,* 2011), using 150 µg of digested and labeled peptides and a Fortis C18 column, 100 mm x 2.1 mm, 5 µm (Fortis Technologies, Marl, Germany). A MALDI TOF/TOF 4800 mass spectrometer (AB SCIEX, Foster City, CA, USA) was used for acquisition and processing of the peptides as described elsewhere (Yakimov *et al.,* 2011). The resulting raw peak lists

of precursors and fragment ions were filtered and exported with the ABI-Extractor tool (Peaks-Bioinformatics Solutions, ON, Canada). Protein identification and quantitation were done with MASCOT v. 2.3.01 (Matrix Science, London, UK) and Phenyx v. 2.6 (GeneBio, Geneva, Switzerland). The search was performed against the predicted protein sequences of *H. tiamatea* SARL4B^T. The concatenated target-decoy database search strategy was used to estimate the false positive rate (below 1%) to improve reliability of the data. The search parameters were as described elsewhere (Yakimov *et al.,* 2011). A minimum of two unique peptides was required for protein identification and further quantification.

Acknowledgments

We thank the crew of the R/V *Urania* for sampling of the Medee DHAL, Jörg Wulf for his help with CARD-FISH analyses, and Rudolf Amann for critical reading of the manuscript. This study was supported by the EU FP7 project MAMBA (FP7-KBBE-2008-226977; http://mamba.bangor.ac.uk/), the Spanish Ministry of Economy and Competitiveness (grant BIO2011-25012) and the Max Planck Society.

References

(2008) List of new names and new combinations previously effectively, but not validly, published. *Int J Syst Evol Microbiol*, **58**: 2471–2472.

Altschul, S.F., Madden, T.L., Schäffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25**: 3389–3402.

Anderson, I., Tindall, B.J., Pomrenke, H., Göker, M., Lapidus, A., Nolan, M.*, et al.* (2009) Complete genome sequence of Halorhabdus utahensis type strain (AX-2T). *Stand Genomic Sci*, **1**: 218–225.

Anderson, I., Scheuner, C., Göker, M., Mavromatis, K., Hooper, S.D., Porat, I.*, et al.* (2011) Novel Insights into the Diversity of Catabolic Metabolism from Ten Haloarchaeal Genomes. *PLoS ONE*, **6**: e20237.

Andrei, A.-Ş., Banciu, H.L., and Oren, A. (2012) Living with salt: metabolic and phylogenetic diversity of archaea inhabiting saline ecosystems. *FEMS Microbiol Lett*, **330**: 1–9.

Antón, J., Rosselló-Mora, R., Rodríguez-Valera, F., and Amann, R. (2000) Extremly Halophilic Bacteria in Crystallizer Ponds from Solar Salterns. *Appl Environ Microbiol*, **66**: 3052–3057.

Antón, J., Llobet-Brossa, E., Rodríguez-Valera, F., and Amann, R. (1999) Fluorescence in situ hybridization analysis of the prokaryotic community inhabiting crystallizer ponds. *Environ Microbiol*, **1**: 517–523.

Antunes, A., Taborda, M., Huber, R., Moissl, C., Nobre, M.F., and da Costa, M.S. (2008a) Halorhabdus tiamatea sp. nov., a non-pigmented, extremely halophilic archaeon from a deep-sea, hypersaline anoxic basin of the Red Sea, and

emended description of the genus Halorhabdus. *Int J Syst Evol Microbiol*, **58**: 215–220.

Antunes, A., França, L., Rainey, F.A., Huber, R., Nobre, M.F., Edwards, K.J., and da Costa, M.S. (2007) Marinobacter salsuginis sp. nov., isolated from the brine-seawater interface of the Shaban Deep, Red Sea. *Int J Syst Evol Microbiol*, **57**: 1035–1040.

Antunes, A., Ngugi, D.K., and Stingl, U. (2011a) Microbiology of the Red Sea (and other) deep-sea anoxic brine lakes. *Environ Microbiol Rep*, **3**: 416–433.

Antunes, A., Alam, I., Bajic, V.B., and Stingl, U. (2011b) Genome Sequence of Halorhabdus tiamatea, the First Archaeon Isolated from a Deep-Sea Anoxic Brine Lake. *J Bacteriol*, **193**: 4553–4554.

Antunes, A., Eder, W., Fareleira, P., Santos, H., and Huber, R. (2003) Salinisphaera shabanensis gen. nov., sp. nov., a novel, moderately halophilic bacterium from the brine-seawater interface of the Shaban Deep, Red Sea. *Extremophiles*, **7**: 29–34.

Antunes, A., Rainey, F.A., Wanner, G., Taborda, M., Pätzold, J., Nobre, M.F., da Costa, M.S., and Huber, R. (2008b) A New Lineage of Halophilic, Wall-Less, Contractile Bacteria from a Brine-Filled Deep of the Red Sea. *J Bacteriol*, **190**: 3580–3587.

Aziz, R.K., Bartels, D., Best, A.A., DeJongh, M., Disz, T., Edwards, R.A., *et al.* (2008) The RAST Server: Rapid Annotations using Subsystems Technology. *BMC Genomics*, **9**: 75.

Baati, H., Guermazi, S., Amdouni, R., Gharsallah, N., Sghi, A., and Ammar, E. (2008) Prokaryotic diversity of a Tunisian multipond solar saltern. *Extremophiles*, **12**: 505–518.

Ball, S.G., and Morell, M.K. (2003) From bacterial glycogen to starch: Understanding the Biogenesis of the Plant Starch Granule. *Annu Rev Plant Biol*, **54**: 207–233.

Bodaker, I., Sharon, I., Suzuki, M.T., Feingersch, R., Shmoish, M., Andreishcheva, E.*, et al.* (2009) Comparative community genomics in the Dead Sea: an increasingly extreme environment. *ISME J*, **4**: 399–407.

Bortoluzzi, G., Borghini, M., La Cono, V., Genovese, L., Foraci, F., Polonia, A.*, et al.* (2011) The Exploration of Deep Hypersaline Anoxic Basins. *Marine research at CNR - Marine Ecology*: 95–108.

Boujelben, I., Gomariz, M., Martínez-García, M., Santos, F., Peña, A., López, C., Antón, J., and Maalej, S. (2012) Spatial and seasonal prokaryotic community dynamics in ponds of increasing salinity of Sfax solar saltern in Tunisia. *Antonie van Leeuwenhoek*, **101**: 845–857.

Brito-Echeverría, J., López-López, A., Yarza, P., Antón, J., and Rosselló-Móra, R. (2009) Occurrence of Halococcus spp. in the nostrils salt glands of the seabird Calonectris diomedea. *Extremophiles*, **13**: 557–565.

Cantarel, B.L., Coutinho, P.M., Rancurel, C., Bernard, T., Lombard, V., and Henrissat, B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*, **37**: D233–238.

Darling, A.E., Mau, B., Perna, N.T., and Stajich, J.E. (2010) progressiveMauve: Multiple Genome Alignment with Gene Gain, Loss and Rearrangement. *PLoS ONE*, **5**: e11147.

De Montalk, G.P., Remaud-Simeon, M., Willemot, R.M., Planchot, V., and Monsan, P. (1999) Sequence Analysis of the Gene Encoding Amylosucrase from Neisseria polysaccharea and Characterization of the Recombinant Enzyme. *J Bacteriol*, **181**: 375–381.

Demergasso, C., Casamayor, E.O., Chong, G., Galleguillos, P., Escudero, L., and Pedrós-Alió, C. (2004) Distribution of prokaryotic genetic diversity in athalassohaline lakes of the Atacama Desert, Northern Chile. *FEMS Microbiol Ecol*, **48**: 57–69.

Eder, W., Jahnke, L.L., Schmidt, M., and Huber, R. (2001) Microbial Diversity of the Brine-Seawater Interface of the Kebrit Deep, Red Sea, Studied via 16S rRNA Gene Sequences and Cultivation Methods. *Appl Environ Microbiol*, **67**: 3077–3085.

Eder, W., Schmidt, M., Koch, M., Garbe-Schönberg, D., and Huber, R. (2002) Prokaryotic phylogenetic diversity and corresponding geochemical data of the brine-seawater interface of the Shaban Deep, Red Sea. *Environ Microbiol*, **4**: 758–763.

Eder, W., Ludwig, W., and Huber, R. (1999) Novel 16S rRNA gene sequences retrieved from highly saline brine sediments of Kebrit Deep, Red Sea. *Arch Microbiol*, **172**: 213–218.

Emanuelsson, O., Brunak, S., Heijne, G. von, and Nielsen, H. (2007) Locating proteins in the cell using TargetP, SignalP and related tools. *Nat Protoc*, **2**: 953–971.

Ewing, B., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using Phred. II. Error Probabilities. *Genome Res*, **8**: 186–194.

Ewing, B., Hiller, L., Wendt, M.C., and Green, P. (1998) Base-Calling of Automated Sequencer Traces Using Phred. I. Accuracy Assement. *Genome Res*, **8**: 175–185.

Fan, H., Xue, Y., Zeng, Y., Zhou, P., and Ma, Y. (2003) Archaeal diversity of Zabuye Lake in Tibet analyzed by culture-independent approach. *Wei Sheng Wu Xue Bao*, **43**: 401–408.

Ferrer, M., Werner, J., Chernikova, T.N., Bargiela, R., Fernández, L., La Cono, V., *et al.* (2012) Unveiling microbial life in the new deep-sea hypersaline Lake Thetis. Part II: a metagenomic study. *Environ Microbiol*, **14**: 268–281.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., *et al.* (2009) The Pfam protein families database. *Nucleic Acids Res*, **38**: D211–222.

Frey, P.A. (1996) The Leloir pathway: a mechanistic imperative for three enzymes to change the stereochemical configuration of a single carbon in galactose. *FASEB J*, **10**: 461–470.

Fuchs, B.M., Glöckner, F.O., Wulf, J., and Amann, R. (2000) Unlabeled Helper Oligonucleotides Increase the In Situ Accessibility to 16S rRNA of Fluorescently Labeled Oligonucleotide Probes. *Appl Environ Microbiol*, **66**: 3603–3607.

Gordon, D. (2003) Viewing and Editing Assembled Sequences Using Consed. *Curr Protoc Bioinformatics*, **2**: 11.2.1–11.2.43.

Grissa, I., Vergnaud, G., and Pourcel, C. (2007) CRISPRFinder: a web tool to identify clustered regularly interspaced short palindromic repeats. *Nucleic Acids Res*, **35**: W52–57.

Han, C.S., and Chain, P. (2006) Finishing Repetitive Regions Automatically with Dupfinisher. In *Proceedings of the 2006 International Conference on Bioinformatics and Computational Biology.* Las Vegas: CSREA Press, pp. 142–147.

Henrissat, B., and Coutinho, P.M. (2001) Classification of glycoside hydrolases and glycosyltransferases from hyperthermophiles. *Method Enzymol*, **330**: 183–201.

Horie, A., Tomita, T., Saiki, A., Kono, H., Taka, H., Mineki, R., *et al.* (2009) Discovery of proteinaceous N-modification in lysine biosynthesis of Thermus thermophilus. *Nat Chem Biol*, **5**: 673–679.

Hsü, K.J., Stoffers, P., and Ross, D.A. (1978) Messinian Evaporites from the Mediterranean and Red Seas. *Mar Geol*, **21**: 71–72.

Hunter, S., Jones, P., Mitchell, A., Apweiler, R., Attwood, T.K., Bateman, A., *et al.* (2011) InterPro in 2011: new developments in the family and domain prediction database. *Nucleic Acids Res*, **40**: D306–312.

Imhoff, J.F., and Rodriguez-Valera, F. (1984) Betaine Is the Main Compatible Solute of Halophilic Eubacteria. *J Bacteriol*, **160**: 478–479.

Itoh, T., Akao, S., Hashimoto, W., Mikami, B., and Murata, K. (2004) Crystal Structure of Unsaturated Glucuronyl Hydrolase, Responsible for the Degradation of Glycosaminoglycan, from Bacillus sp. GL1 at 1.8 A Resolution. *J Biol Chem*, **279**: 31804–31812.

Jakobsen, T.F., Kjeldsen, K.U., and Ingvorsen, K. (2006) Desulfohalobium utahense sp. nov., a moderately halophilic, sulfate-reducing bacterium isolated from Great Salt Lake. *Int J Syst Evol Microbiol*, **56**: 2063–2069.

Jiang, H., Dong, H., Yu, B., Liu, X., Li, Y., Ji, S., and Zhang, C.L. (2007) Microbial response to salinity change in Lake Chaka, a hypersaline lake on Tibetan plateau. *Environ Microbiol*, **9**: 2603–2621.

Jiao, N., Feng, F., and Wei, B. (2006) Proteorhodopsin—A new path for biological utilization of light energy in the sea. *Chinese Sci Bull*, **51**: 889–896.

Kamat, S.S., Williams, H.J., and Raushel, F.M. (2011) Intermediates in the transformation of phosphonates to phosphate by bacteria. *Nature*, **480**: 570–573.

Kanehisa, M., and Goto, S. (2000) KEGG: Kyoto Encyclopedia of Genes and Genomes. *Nucleic Acids Res*, **28**: 27–30.

Kather, B., Stingl, K., van der Rest, M.E., Altendorf, K., and Molenaar, D. (2000) Another Unusual Type of Citric Acid Cycle Enzyme in Helicobacter pylori: the Malate:Quinone Oxidoreductase. *J Bacteriol*, **182**: 3204–3209.

Khotimchenko, Y., Khozhaenko, E., Kovalev, V., and Khotimchenko, M. (2012) Cerium Binding Activity of Pectins Isolated from the Seagrasses Zostera marina and Phyllospadix iwatensis. *Mar Drugs*, **10**: 834–848.

Kitaoka, M., Matsuoka, Y., Mori, K., Nishimoto, M., and Hayashi, K. (2012) Characterization of a Bacterial Laminaribiose Phosphorylase. *Biosci Biotechnol Biochem*, **76**: 343–348.

Kjeldsen, K.U., Loy, A., Jakobsen, T.F., Thomsen, T.R., Wagner, M., and Ingvorsen, K. (2007) Diversity of sulfate-reducing bacteria from an extreme hypersaline sediment, Great Salt Lake (Utah). *FEMS Microbiol Ecol*, **60**: 287–298.

Klindworth, A., Pruesse, E., Schweer, T., Peplies, J., Quast, C., Horn, M., *et al.* (2012) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next-generation sequencing-based diversity studies. *Nucleic Acids Res*, **41**: e1.

Krogh, A., Larsson, B., Heijne, G. von, and Sonnhammer, E.L. (2001) Predicting transmembrane protein topology with a hidden markov model: application to complete genomes. *J Mol Biol*, **305**: 567–580.

Legault, B.A., Lopez-Lopez, A., Alba-Casado, J., Doolittle, W.F., Bolhuis, H., Rodriguez-Valera, F., and Papke, R.T. (2006) Environmental genomics of "Haloquadratum walsbyi" in a saltern crystallizer indicates a large pool of accessory genes in an otherwise coherent species. *BMC Genomics*, **7**: 171.

Lowe, T.M., and Eddy, S.R. (1997) tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. *Nucleic Acids Res*, **25**: 955–964.

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

Ma, K., Weiss, R., and Adams, M.W.W. (2000) Characterization of Hydrogenase II from the Hyperthermophilic Archaeon Pyrococcus furiosus and Assessment of Its Role in Sulfur Reduction. *J Bacteriol*, **182**: 1864–1871.

Ma, K., Schicho, R.N., Kelly, R.M., and Adams, M.W.W. (1993) Hydrogenase of the hyperthermophile Pyrococcus furiosus is an elemental sulfur reductase or sulfhydrogenase: Evidence for a sulfur-reducing hydrogenase ancestor. *Proc Natl Acad Sci USA*, **90**: 5341–5344.

Mai, X., and Adams, M.W.W. (1996) Characterization of a Fourth Type of 2-Keto Acid-Oxidizing Enzyme from a Hyperthermophilic Archaeon: 2-Ketoglutarate Ferredoxin Oxidoreductase from Thermococcus litoralis. *J Bacteriol*, **178**: 5890–5896.

Makhdoumi-Kakhki, A., Amoozegar, M.A., Kazemi, B., PaiC, L., and Ventosa, A. (2012) Prokaryotic Diversity in Aran-Bidgol Salt Lake, the Largest Hypersaline Playa in Iran. *Microbes Environ*, **27**: 87–93.

Malykh, A., Malykh, O., Polushin, N., Kozyavkin, S., and Slesarev, A. (2004) Finishing "Working Draft" BAC Projects by Directed Sequencing With ThermoFidelase and Fimers. *Method Mol Biol*, **255**: 295–308.

Martens, E.C., Lowe, E.C., Chiang, H., Pudlo, N.A., Wu, M., McNulty, N.P., *et al.* (2011) Recognition and Degradation of Plant Cell Wall Polysaccharides by Two Human Gut Symbionts. *PLoS Biol*, **9**: e1001221.

Maturrano, L., Santos, F., Rossello-Mora, R., and Anton, J. (2006) Microbial Diversity in Maras Salterns, a Hypersaline Environment in the Peruvian Andes. *Appl Environ Microbiol*, **72**: 3887–3895.

Meyer, F., Goesmann, A., McHardy, A.C., Bartels, D., Bekel, T., Clausen, J.*, et al.* (2003) GenDB--an open source genome annotation system for prokaryote genomes. *Nucleic Acids Res*, **31**: 2187–2195.

Murray, A.E., Kenig, F., Fritsen, C.H., McKay, C.P., Cawley, K.M., Edwards, R.*, et al.* (2012) Microbial life at -13 °C in the brine of an ice-sealed Antarctic lake. *Proc Natl Acad Sci USA*, **109**: 20626–20631.

Mutlu, M.B., Martínez-Garcí-a, M., Santos, F., Peña, A., Guven, K., and Antón, J. (2008) Prokaryotic diversity in Tuz Lake, a hypersaline environment in Inland Turkey. *FEMS Microbiol Ecol*, **65**: 474–483.

Nelson-Sathi, S., Dagan, T., Landan, G., Janssen, A., Steel, M., McInerney, J.O., Deppenmeier, U., and Martin, W.F. (2012) Acquisition of 1,000 eubacterial genes physiologically transformed a methanogen at the origin of Haloarchaea. *Proc Natl Acad Sci USA*, **109**: 20537–20542.

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

Ning, Z., Cox, A.J., and Mullikin, J.C. (2001) SSAHA: A Fast Search Method for Large DNA Databases. *Genome Res*, **11**: 1725–1729.

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

Nolla-Ardèvol, V., Strous, M., Sorokin, D.Y., Merkel, A.Y., and Tegetmeyer, H.E. (2012) Activity and diversity of haloalkaliphilic methanogens in Central Asian soda lakes. *J Biotechnol*, **161**: 167–173.

O'Malley, M.A. (2008) 'Everything is everywhere: but the environment selects': ubiquitous distribution and ecological determinism in microbial biogeography. *Stud Hist Philos Sci Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, **39**: 314–325.

Okada, G., and Hehre, E.J. (1974) New Studies on Amylosucrase, a Bacterial a-D-Glucosylase That Directly Converts Sucrose to a Glycogen-like a-Glucan. *J Biol Chem*, **249**: 126–135.

Oren, A. (1991) Anaerobic growth of halophilic archaeobacteria by reduction of fumarate. *J Gen Microbiol*, **137**: 1387–1390.

Oren, A. (2008) Microbial life at high salt concentrations: phylogenetic and metabolic diversity. *Saline Systems*, **4**: 2.

Oren, A., and Gurevich, P. (1994) Production of D-lactate, acetate, and pyruvate from glycerol in communities of halophilic archaea in the Dead Sea and in saltern crystallizer ponds. *FEMS Microbiol Ecol*, **14**: 147–155.

Pan, H.-l., Zhou, C., Wang, H.-l., Xue, Y.-f., and Ma, Y.-h. (2006) Diversity of Halophilic Archaea in Hypersaline lakes of Inner Mongolia, China. *Wei Sheng Wu Xue Bao*, **46**: 1–6.

Park, H.-W., Kim, S.-T., Sancar, A., and Deisenhofer, J. (1995) Crystal Structure of DNA Photolyase from Escherichia coli. *Science*, **268**: 1866–1872.

Pašić, L., Ulrih, N.P., Črnigoj, M., Grabnar, M., and Velikonja, B.H. (2007) Haloarchaeal communities in the crystallizers of two adriatic solar salterns. *Can J Microbiol*, **53**: 8–18.

Pautot, G., Guennoc, P., Coutelle, A., and Lyberis, N. (1984) Discovery of a large brine deep in the northern Red Sea. *Nature*, **310**: 133–136.

Pernthaler, A., Pernthaler, J., and Amann, R. (2004) Sensitive multi-color fluorescence in situ hybridization for the identification of environmental microorganisms. In *Molecular Microbial Ecology Manual.* Kowalchuk, G.A., de Bruijn, F.J., Head, I.M., Akkermans, A.D.L., and van Elsas, J.D. (eds) : Springer, pp. 2613–2627.

Pernthaler, A., Pernthaler, J., and Amann, R. (2002) Fluorescence In Situ Hybridization and Catalyzed Reporter Deposition for the Identification of Marine Bacteria. *Appl Environ Microbiol*, **68**: 3094–3101.

Pilcher, R.S., and Blumstein, R.D. (2007) Brine volume and salt dissolution rates in Orca Basin, northeast Gulf of Mexico. *AAPG Bulletin*, **91**: 823–833.

Podzelinska, K., He, S.-M., Wathier, M., Yakunin, A., Proudfoot, M., Hove-Jensen, B., Zechel, D.L., and Jia, Z. (2009) Structure of PhnP, a Phosphodiesterase of the Carbon-Phosphorus Lyase Pathway for Phosphonate Degradation. *J Biol Chem*, **284**: 17216–17226.

Popper, Z.A., Michel, G., Hervé, C., Domozych, D.S., Willats, W.G., Tuohy, M.G., Kloareg, B., and Stengel, D.B. (2011) Evolution and Diversity of Plant Cell Walls: From Algae to Flowering Plants. *Annu Rev Plant Biol*, **62**: 567–590.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2004) NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. *Nucleic Acids Res*, **33**: D501–504.

Quast, C., Pruesse, E., Yilmaz, P., Gerken, J., Schweer, T., Yarza, P., Peplies, J., and Glockner, F.O. (2012) The SILVA ribosomal RNA gene database project:

improved data processing and web-based tools. *Nucleic Acids Res*, **41**: D590–596.

Rawls, K.S., Martin, J.H., and Maupin-Furlow, J.A. (2011) Activity and Transcriptional Regulation of Bacterial Protein-Like Glycerol-3-Phosphate Dehydrogenase of the Haloarchaea in Haloferax volcanii. *J Bacteriol*, **193**: 4469–4476.

Remelli, W., Guerrieri, N., Klodmann, J., Papenbrock, J., Pagani, S., Forlani, F., and Pastore, A. (2012) Involvement of the Azotobacter vinelandii Rhodanese-Like Protein RhdA in the Glutathione Regeneration Pathway. *PLoS ONE*, **7**: e45193.

Rice, P., Longden, I., and Bleasby, A. (2000) EMBOSS: The European Molecular Biology Open Software Suite. *Trends Genet*, **16**: 276–277.

Richter, M., Lombardot, T., Kostadinov, I., Kottmann, R., Duhaime, M., Peplies, J., and Glöckner, F. (2008) JCoast – A biologist-centric software tool for data mining and comparison of prokaryotic (meta)genomes. *BMC Bioinformatics*, **9**: 177.

Sakai, S., Takaki, Y., Shimamura, S., Sekine, M., Tajima, T., Kosugi, H., *et al.* (2011) Genome Sequence of a Mesophilic Hydrogenotrophic Methanogen Methanocella paludicola, the First Cultivated Representative of the Order Methanocellales. *PLoS ONE*, **6**: e22898.

Sakon, J., Irwin, D., Wilson, D.B., and Karplus, P.A. (1997) Structure and mechanism of endo/exocellulase E4 from Thermomonospora fusca. *Nat Struct Biol*, **4**: 810–818.

Sangwan, N., Lata, P., Dwivedi, V., Singh, A., Niharika, N., Kaur, J., *et al.* (2012) Comparative Metagenomic Analysis of Soil Microbial Communities across Three Hexachlorocyclohexane Contamination Levels. *PLoS ONE*, **7**: e46219.

Schryvers, A., and Weiner, J.H. (1981) The Anaerobic sn-Glycerol-3-phosphate Dehydrogenase of Escherichia coli. *J Biol Chem*, **256**: 9959–9965.

Stahl, D.A., and Amann, R. (1991) Development and application of nucleic acid probes. In *Nucleic acid techniques in bacterial systematics.* Stackebrandt, E., and Goodfellow, M. (eds) . Chichester, UK: John Wiley and Sons, pp. 205–248.

Stamatakis, A., Ludwig, T., and Meier, H. (2005) RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. *Bioinformatics*, **21**: 456–463.

Tatusov, R.L., Fedorova, N.D., Jackson, J.D., Jacobs, A.R., Kiryutin, B., Koonin, E.V.*, et al.* (2003) The COG database: an updated version includes eukaryotes. *BMC Bioinformatics*, **4**: 41.

Teeling, H., Waldmann, J., Lombardot, T., Bauer, M., and Glöckner, F. (2004) TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *BMC Bioinformatics*, **5**: 163.

Teeling, H., Fuchs, B.M., Becher, D., Klockow, C., Gardebrecht, A., Bennke, C.M.*, et al.* (2012) Substrate-Controlled Succession of Marine Bacterioplankton Populations Induced by a Phytoplankton Bloom. *Science*, **336**: 608–611.

The UniProt Consortium (2010) Ongoing and future developments at the Universal Protein Resource. *Nucleic Acids Res*, **39**: D214–219.

Thiele, S., Fuchs, B.M., and Amann, R. (2011) Identification of Microorganisms Using the Ribosomal RNA Approach and Fluorescence In Situ Hybridization. In *Treatise on Water Science.* Wilderer, P. (ed). Oxford, United Kingdom: Elsevier, pp. 171–189.

Tsiamis, G., Katsaveli, K., Ntougias, S., Kyrpides, N., Andersen, G., Piceno, Y., and Bourtzis, K. (2008) Prokaryotic community profiles at different operational stages of a Greek solar saltern. *Res Microbiol*, **159**: 609–627.

Urakawa, H., Martens-Habbena, W., and Stahl, D.A. (2010) High Abundance of Ammonia-Oxidizing Archaea in Coastal Waters, Determined Using a Modified DNA Extraction Method. *Appl Environ Microbiol*, **76**: 2129–2135.

van der Wielen, P.W.J.J., Bolhuis, H., Borin, S., Daffonchio, D., Corselli, C., Giuliano, L., *et al.* (2005) The Enigma of Prokaryotic Life in Deep Hypersaline Anoxic Basins. *Science*, **307**: 121–123.

Wainø, M., and Ingvorsen, K. (2003) Production of beta-xylanase and beta-xylosidase by the extremely halophilic archaeon Halorhabdus utahensis. *Extremophiles*, **7**: 87–93.

Wainø, M., Tindall, B.J., and Ingvorsen, K. (2000) Halorhabdus utahensis gen. nov., sp. nov., an aerobic, extremely halophilic member of the Archaea from Great Salt Lake, Utah. *Int J Syst Evol Microbiol*, **50**: 183–190.

Wallner, G., Amann, R., and Beisker, W. (1993) Optimizing fluorescent in situ hybridization with rRNA-targeted oligonucleotide probes for flow cytometric identification of microorganisms. *Cytometry*, **14**: 136–143.

Yakimov, M.M., La Cono, V., Smedile, F., DeLuca, T.H., Juárez, S., Ciordia, S., *et al.* (2011) Contribution of crenarchaeal autotrophic ammonia oxidizers to the dark primary production in Tyrrhenian deep waters (Central Mediterranean Sea). *ISME J*, **5**: 945–961.

Yernool, D.A., McCarthy, J.K., Eveleigh, D.E., and Bok, J.-D. (2000) Cloning and Characterization of the Glucooligosaccharide Catabolic Pathway beta -Glucan Glucohydrolase and Cellobiose Phosphorylase in the Marine Hyperthermophile Thermotoga neapolitana. *J Bacteriol*, **182**: 5172–5179.

Zaporozhets, T.S. (2003) Neutrophil activation by sea hydrobiont biopolymers. *Antibiot Khimioter*, **48**: 3–7.

Figure legends

Figure 1: Maximum likelihood tree of the family *Halobacteriaceae.*

The tree was calculated with RAxML v. 7.0.3 (Stamatakis *et al.,* 2005) with *Methanospirillum hungatei* JF-1 as outgroup. The scale bar represents 10% estimated sequence divergence.

Figure 2: Circular representation of the chromosome (A) and plasmid (B) of *H. tiamatea* SARL4B^T.

From inside to outside: GC content, GC skew, DNA curvature, DNA bending, deviation from the average tetranucleotide composition, CAZymes (blue: glycoside hydrolase, red: glycosyl transferase, green: carbohydrate esterase, orange: polysaccharide lyase, cyan: carbohydrate binding module), RNAs (red: rRNA, green: tRNA, orange: other RNA), genes in reverse direction, genes in forward direction. GC content and GC skew were calculated with a self-written PERL script (sliding windows: 5 kbp for chromosome; 0.5 kbp for plasmid). DNA curvature and bending were calculated with the program banana from the EMBOSS package (Rice *et al.,* 2000). TETRA (Teeling *et al.,* 2004) was used for the calculation of the deviation from the average tetranucleotide composition (sliding windows: 5 kbp for chromosome; 1 kbp for plasmid).

Figure 3: Whole genome alignment between *H. utahensis* AX-2^T and *H. tiamatea* SARL4B^T.

A: Mapping of the chromosomes based on SSAHA2 v. 2.5 (Ning *et al.,* 2001). The genome of *H. tiamatea* SARL4B^T was split into 50 bp fragments and subsequently mapped on the *H. utahensis* AX-2^T. B: Mauve (v. 2.3.1) (Darling *et al.,* 2010) whole

genome alignment plot. The diagrams inside the boxes represent the similarity in every genomic area by comparing the two genomes with a bidirectional BLASTn analysis.

Figure 4: Identification of a *Halorhabdus*-species in a sample from the eastern Mediterranean DHAL Medee.

A: DAPI staining; B: hybridization with the probe Halo178 and helper probes Halo178-h1 and Halo178-h2, showing typical pleomorphic cells of *Halorhabdus* species. Images were post-processed with Autoquant X (A and B) and Imaris 7.4.0 (only B).

Tables

Table 1: General characteristics of the *H. utahensis* AX-2^T and *H. tiamatea* SARL4B^T genomes.

	H. utahensis AX-2^T	*H. tiamatea* SARL4B^T
Contigs	1 chromosome	1 chromosome, 1 plasmid
chromosome size	3,116,795 bp	2,815,791 bp
plasmid size	–	330,369 bp
GC content	62.9%	62.7%
total genes	2,998	3,023
CDS	2,687,322 bp	2,616,558 bp
coding density	86.2%	83.2%
genes with functions	2,243 (74.8%)	1,974 (65.3%)
hypothetical proteins	736 (24.5%)	480 (15.9%)
conserved hypothetical proteins	19 (0.6%)	569 (18.8%)
rRNAs	3	3
tRNAs	45	46

Table 2: CAZymes in the genomes of *H. utahensis* AX-2^T (according to the CAZy database Cantarel *et al.,* 2009 as of 2012/11/10), *H. tiamatea* SARL4B^T, and the ANR26 enrichment from the eastern Mediterranean DHAL Medee.

	H. utahensis AX-2^T	*H. tiamatea* SARL4B^T	ANR26 enrichment
GH2	4 (1.28)	4 (1.27)	4 (1.02)
GH3	7 (2.25)	6 (1.91)	5 (1.28)
GH4	2 (0.64)	1 (0.32)	0 (0.00)
GH5	7 (2.25)	2 (0.64)	2 (0.51)
GH9	1 (0.32)	1 (0.32)	0 (0.00)
GH10	4 (1.28)	7 (2.22)	3 (0.77)
GH11	2 (0.64)	0 (0.00)	0 (0.00)
GH13	1 (0.32)	7 (2.22)	9 (2.30)
GH31	0 (0.00)	1 (0.32)	1 (0.26)
GH32	0 (0.00)	2 (0.64)	1 (0.26)
GH42	0 (0.00)	1 (0.32)	0 (0.00)
GH43	4 (1.28)	3 (0.95)	2 (0.51)
GH51	1 (0.32)	6 (1.91)	0 (0.00)
GH67	1 (0.32)	1 (0.32)	0 (0.00)
GH77	1 (0.32)	1 (0.32)	1 (0.26)
GH88	0 (0.00)	1 (0.32)	0 (0.00)
GH93	0 (0.00)	1 (0.32)	0 (0.00)
GH94	2 (0.64)	3 (0.95)	3 (0.77)
GH95	1 (0.32)	1 (0.32)	0 (0.00)
GH97	1 (0.32)	1 (0.32)	0 (0.00)

The first number represents absolute counts, and the number in parentheses the relative number per Mbp.

Supplementary material

The following supplementary material is available for this article online:

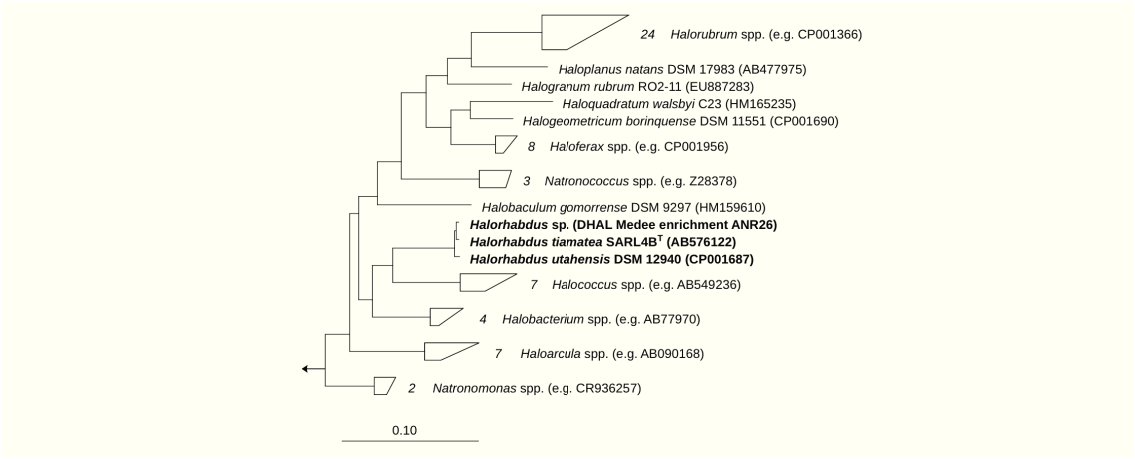
Supplementary text. Transporters; Motility and chemotaxis; Fermentation products; Biotechnological aspects; Biogeographical aspects

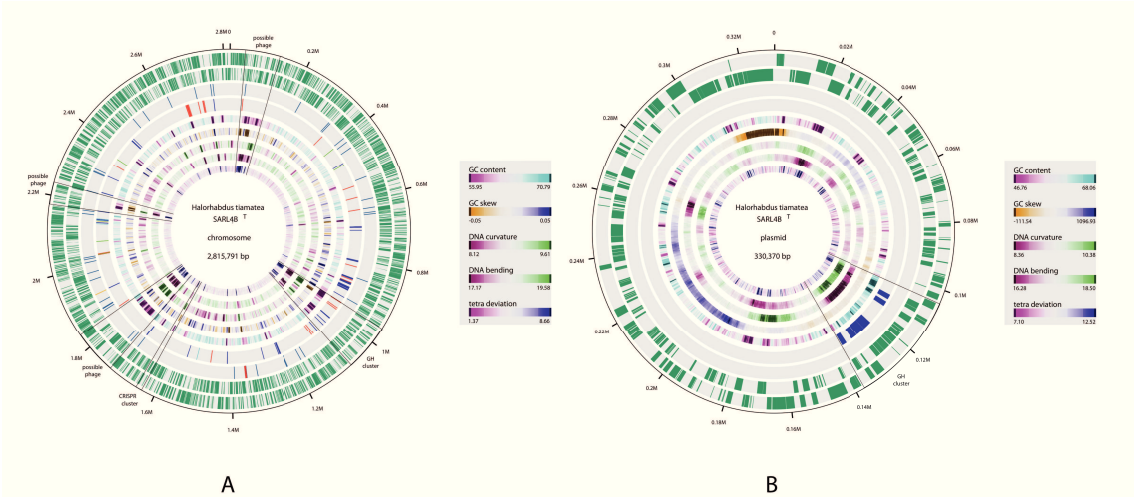
Figure S1. Phylogenetic tree of known archaeal GH13 family genes and GH13 genes of *H. tiamatea* SARL4B^T

Figure S2. Numbers of glycoside hydrolases in CAZyme-rich *Archaea*.

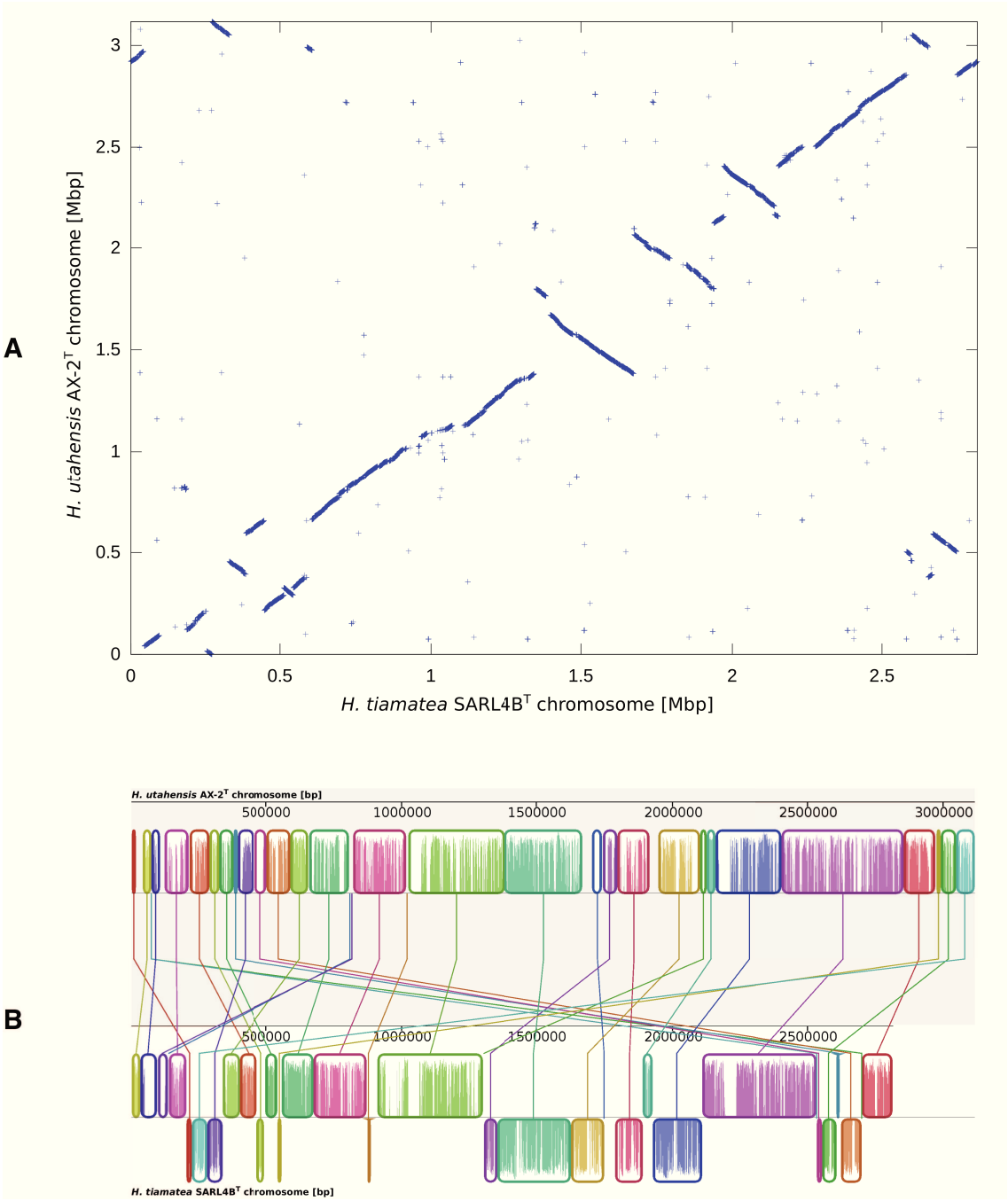
Table S1. Glycoside hydrolases in *H. tiamatea* SARL4B^T.

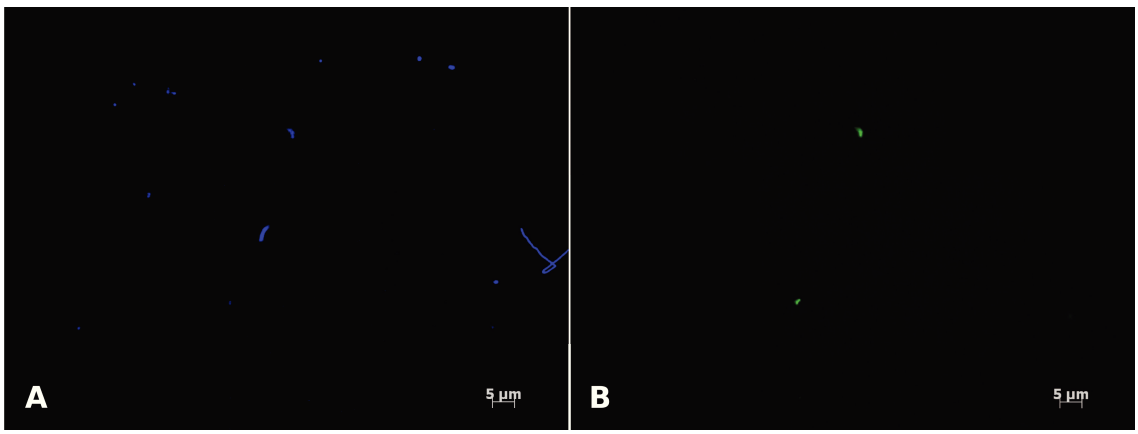
Table S2. Proteome analysis of oxygen stress response


24 Halorubrum spp. (e.g. CP001366)
Haloplanus natans DSM 17983 (AB477975)
Halogranum rubrum RO2-11 (EU887283)
Haloquadratum walsbyi C23 (HM165235)
Halogeometricum borinquense DSM 11551 (CP001690)
8 Haloferax spp. (e.g. CP001956)
3 Natrinococcus spp. (e.g. Z28378)
Halobaculum gomorrense DSM 9297 (HM159610)
Halorhabdus sp. (DHAL Medee enrichment ANR26)
Halorhabdus tiamatea SARL4B T (AB576122)
Halorhabdus utahensis DSM 12940 (CP001687)
7 Halococcus spp. (e.g. AB549236)
4 Halobacterium spp. (e.g. AB77970)
7 Haloarcula spp. (e.g. AB090168)
2 Natronomonas spp. (e.g. CR936257)
0.10


Halorhabdus tiamatea SARL4B T
chromosome
2,815,791 bp
GC content
GC skew
DNA curvature
DNA bending
tetra deviation
Halorhabdus tiamatea SARL4B T
plasmid
330,370 bp
GC content
GC skew
DNA curvature
DNA bending
tetra deviation

A B


A
H. utahensis AX-2T chromosome [Mbp]
H. tiamatea SARL4BT chromosome [Mbp]
0
0.5
1
1.5
2
2.5
3
B
H. utahensis AX-2T chromosome [bp]
500000
1000000
1500000
2000000
2500000
3000000
H. tiamatea SARL4BT chromosome [bp]


A
5 µm
B
5 µm

3. Discussion

The explanatory power of CAZyme profiles depends on the accurate identification of CAZyme genes. Our incomplete knowledge leads to imperfect classification of CAZyme genes in software pipelines. On the one hand, we fail to recognize unknown new families. On the other hand, we tend to consistently over-estimate some families because of their mis-annotation in databases. Before a discussion of the CAZyme results, it is hence important to address these methodical issues so that we can assess the margin of error in the results. In addition, understanding the different causes of mis-identification is the first step towards improvements of software pipelines such as Trident. Therefore, the first chapter of this discussion is dedicated to these methodological issues.

The breakdown of polysaccharides represents only a subset of the macromolecule turnover in a microbial community. CAZymes represent only a subset of all carbohydrate-related enzymes. Sugar transporters, sulfatases and carbohydrate-binding proteins are necessary supplements to CAZymes. They deserve our attention and will be discussed in chapter 3.2. Chapter 3.3 begins with a quick recap of the major findings of the MIMAS study and then discusses the microbial interactions based on CAZymes. Inspired by the conclusion drawn in chapter 3.3, chapter 3.4 extracts the CAZyme results of Ulvibacter , *F*ormosa and Polaribacter from the MIMAS study and compares them with eight other *F*lavobacteriaceae genomes. Based on the comparisons, it models the niche-adaptation of *F*lavobacteriaceae and discusses their associations with diatoms and brown algae. Chapter 3.5 focuses on the metagenomes of Hot Lake and Logatchev. As summarized in Chapter 3.6, these three studies together with the genome characterizations of *H*alorhabdus tiamatea SARL4B^T and *F*ormosa agariphila KMM 3901^T have demonstrated CAZyme profiling as a useful tool in genomic studies.

3.1 The accuracy of CAZyme identification

One of the design philosophies of the Trident software pipeline is to identify all CAZymes in a given genome by combining three different approaches. However, such an approach is vulnerable to false positives. If we define the ratio of false postives as P(positive|nonCAZymes) and similarly, the ratio of true positive CAZymes as P(positive|CAZymes), then according to the Bayes' theorem, the probability that a hit is a false positive can be calculated as follows:

$$P(\text{nonCAZymes}|\text{positive}) = \frac{P(\text{nonCAZymes}) \cdot P(\text{positive}|\text{nonCAZymes})}{P(\text{nonCAZymes}) \cdot P(\text{positive}|\text{nonCAZymes}) + P(\text{CAZymes}) \cdot P(\text{positive}|\text{CAZymes})}$$

If we assume P(nonCAZymes)=98%, P(CAZymes)=2% and very optimistically P(positive|CAZymes)=100% and P(positive|nonCAZymes)=5%, then the false positive ratio is at a surprisingly high level of 71%. Using more realistic numbers this could well mean that in practice about three fourth of reported CAZymes are in fact non-CAZyme false positives. This phenomenon has already been noticed in earlier studies of land plant CAZymes [3]. It is due to the low CAZyme frequencies in the genomes (about 1-2% [1]). Even if all CAZymes would be identified with 100% accuracy, their numbers would be small. In contrast, even if small portions of the non-CAZyme genes were falsely identified, their numbers would still be substantial.

There are at least three causes for false positives [3]. First, some query sequences align well to non-catalytic parts of the reference sequences. Since the algorithm does not distinguish catalytic and non-catalytic sites, it will interpret these sequences as close homologues. Second, there are reference genes that are in fact non-CAZymes but are annotated as such because they contain domains similar to functional CAZymes. The unknown query sequences will also be annotated as CAZymes when they match these pseudo-CAZymes. One such example is the chrk1 gene of tobacco, which is related to chitinases, but has lost its chitinase catalytic domain and therefore this enzymatic activity [3]. The third reason is subtle. Enzymes like soybean hydroxyisourate hydrolase have kept both, the catalytic residues and the catalytic mechanism of a □-glucosidase, yet they catalyze reactions on a non-carbohydrate substrate [3].

However, these three kinds of false positives should be treated differently. In the first case, false positives can be reduced by careful examination of the alignments. In fact, a current follow-up project is to achieve just that. The plan will be discussed in more detail in the outlook. In comparison, the second and the third type of wrongful hits require much more effort to correct. The wrong CAZyme annotations need to be corrected in the reference database. This calls for experiments to elucidate their real substrate specificities. However, the biggest challenge is how to identify those pseudo-CAZymes in the database in the first place. It is hard to estimate how severe these two types of errors are in the CAZy database. One possible way to detect these errors is comparing results from different databases with different search algorithms. Moreover, this approach is considered to be too crude to sensitively identify some highly similar pseudo-CAZymes such as the soybean hydroxyisourate hydrolase. A more sophisticated and accurate approach would be molecular modeling, which could

predict the enzyme-substrate interaction to some extent, but this is currently too time-consuming and technically demanding for large-scale sequencing projects.

3.2 The necessity of characterizing carbohydrate metabolism genes other than CAZymes

Although this thesis focuses mainly on CAZymes, the profiling of CAZymes alone is by no means sufficient to understand the lifestyle of a microorganism with respect to carbohydrates. Sulfatases, the components of TonB-dependent transport systems and ABC transporters were for example part of the discussion in the MIMAS study (chapter 2.2). Also, the distributions of peptidoglycan-binding domains in the Logatchev study revealed insightful patterns of prokaryotic protein glycosylation.

The first obvious reason that CAZymes alone are not enough is that carbohydrate degradation involves more enzymes than solely CAZymes. As mentioned in the introduction, marine carbohydrates are often decorated with negatively charged moieties, in particular sulfate groups. These sulfate groups have to be removed by sulfatases before the rest of the carbohydrate molecule can be cleaved and funneled into the monosaccharide catabolic pathways. Also, membrane transporters play a central role in microbial carbohydrate uptake [119]. Once the carbohydrate substrates are in the vicinity of CAZymes, the enzymes need certain mechanisms to recognize and bind these substrates. Some of the respective binding domains are better formulated in Pfam than in CAZy, as for example the LysM domain [120].

The second reason is that CAZymes are not perfectly correlated with the other carbohydrate-related genes, even though some of them are functionally connected. For example, the susD-like and GH16 gene frequencies in the flavobacterial taxobins and reference genomes did not regress well linearly within the MIMAS study (R^2=0.31, N=18), nor did the regression between sulfatase and GH16 genes (R^2=0.48, N=18). Such poor correlations between ABC transporters, TonB-dependent transporters and trophic strategies have also been observed in a previous study by Tang □□ □□ [119]. Tang □□ □□ even pointed out that the frequencies of ABC transporters and TonB-dependent transporters were negatively correlated. The reasons for such poor correlations can be multifold. First, gene frequencies describe the presence of genes in a genome or metagenome, but not their expression. In other words, gene frequencies are features of the genotype, not of the phenotype. It is the expression of genes that dictate different trophic strategies. The actual expression of the genes and their effects on the organisms may differ greatly from what the gene copy numbers may indicate. A gene's copy number is just one of the factors that influences its expression level. Others include promoters, regulatory elements, RNAi to name just a few. Even when

genes are expressed, auxiliary proteins, cofactors and substrates can have an impact on the effect and the range of the products. Therefore, it is a long way between gene frequencies and expressions. A genome's gene frequency pattern is only a blueprint of an organism's lifestyle(s) but not the lifestyle itself. Thus CAZymes are just one facet of the metabolism. We also need additional information such as transporters and sulfatases to get a more complete picture, and, even more importantly, gene expression data for confirmation.

3.3 CAZymes provided insight into the carbohydrate flow in the bacterioplankton bloom in 2009

The relatively low bacterial biodiversity after the phytoplankton bloom in 2009 concealed the complex trophic relations underneath. Although highly sensitive measurement methods and high-throughput sequencing techniques were used, the information revealed were still far from complete, as for example polysaccharides were not measured [1]. After the phytoplankton bloom, there were at least three waves of bacterioplankton blooms. The first was represented by the genus [illegible]. Afterward, there were members of *F*[illegible]. They appeared about two weeks after the chlorophyll peak at 23.03.2009. Since the silica concentrations in the water samples also started to rise at the same time, it is reasonable to conclude that the *F*[illegible] bloom marked the beginning of the diatom downfall. However, the *F*[illegible] bloom was short-lived as it lasted merely about one week. Members of the genus *F*[illegible] were subsequently replaced by two other bacterial genera [illegible] and [illegible]. They both peaked at 14.04.2009. Afterwards, [illegible] disappeared quickly. In comparison, the [illegible] bloom persisted one week longer with a second peak.

A diverse microbial community could not have become established without a vast molecule exchange network. Polymers such as nucleic acids, proteins, lipids and carbohydrates are all intermediates in this network. They can be synthesized, converted, degraded or recycled by nucleases, proteases, lipases and CAZymes. The concentrations of these intermediates are determined by the abundances and productivities of the organisms, which are in turn under the influence of various abiotic factors such as light, temperature and salinity. All these factors contribute to the dynamics of the community. Although the direct interaction between two factors can be straightforward, the whole system is notoriously hard to understand. In fact, all its constituents can have so many possible interactions and variations that the outcomes are difficult to predict. This is a so-called "complex system" [121]. In such a system, the components are highly connected and changes have cascade and multiplying effects. In other words, even if the initial status is known, a small parameter change can bring

the system into a completely different state. For this reason, it is formidably hard to either predict the system based on initial parameters or break it down into the initial state. A diverse microbial community has all the characteristics of a complex system.

A case in point was the microbial succession in spring 2009 in the North Sea. Different CAZyme, protease, lipase, sulfatase and transporter profiles of the occurring bacterial genera indicated that they interacted differently with various substrates [15]. For example, the relatively strong expressions of ABC and TRAP transporters in □ □□□□□□□□ sp. at 2009.04.14 and 2009.04.21 emphasized their reliance on peptides, phosphate, monosaccharides and other monomers [15]. The SAR92 clade had a clear degradation potential for laminarin. The ABC transporter profiling showed that the □ □□□□□□□□□□□clade had the ability of taking up carbohydrates, while the SAR11 clade preferred nitrogen-containing DOC [122]. Even if the abiotic factors were ignored, these variations in nutrient preferences of different bacterial clades alone rendered the whole event intractable. Despite the complexity, this thesis tries to prove that it was still possible to understand a specific route of carbohydrate metabolisms.

As mentioned before, carbohydrates are only one class of the intermediates in the community. The substrates and bacterial clades discussed in this thesis were again only samples from the whole population. Nevertheless, it was not a random choice to characterize this specific route of carbohydrate metabolism. First, the biodiversity from 31.03.2009 to 14.04.2009 was relatively low for the communities during the succession were dominated by well characterized clades with distinct CAZyme profiles [15]. They were the key players. Second, compared to proteases and lipases, the CAZymes responsible for the turnover were relatively well characterized. Third, these carbohydrates were not only organism-specific, but also important for the growth of the diatoms, *F*□□□ □□□ spp., □ □□□□□□□ spp. and □ □□□□□□□□□□□□ spp.. For example, chrysolaminarins constitute at least 10-50% of the cellular carbon in diatoms [10, 40, 41]. This carbohydrate flow could reveal the general trophic connections among the key players without too many confounding details. It was the first step towards our understanding of the succession mechanism.

Based on the CAZyme profiles of *F*□□□ □□□ from the reference genomes and metagenomes, I hypothesize that *F*□□□ □□□ acted as primary diatom degrader. The first indication for this ecological role is the high sulfatase frequencies in the *F*□□□ □□□ spp. genomes. Although the chemical structures of diatom TEP are not clear, it is known that TEP is sulfated [38]. Therefore, *F*□□□ □□□ spp. was likely able to disintegrate the extracellular TEP and thereby effectively isolating the diatoms. Once the diatom cells were accessible, *F*□□□ □□□ cells could begin to degrade the gaskets between the frustules. The gasket consists of callose, an insoluble □-1,3-glucan synthesized by

GT48 □-1,3-glucan synthases. Members of GT48 are only found in eukaryotes and there are three copies of GT48 in the genome of the diatom Thalassiosira pseudonana CCMP1335.

Both callose and chrysolaminarin contain □-1,3-linked glucan chains. Therefore they can be both degraded by □-1,3-glucanases. Endo-glucanase activity is spread over several CAZyme families including GH16 and GH17. Exo-glucanases can be found in GH5. Both types of CAZymes were enriched in all three *F*ormosa spp. reference genomes and the *F*ormosa taxobin at 07.04.2009. Within those genomes and taxobin, all GH16 genes contained the signature catalytic triad □E-D-E″ [29, 123]. After the callose was removed, the diatom frustules were ready to be opened. However, chrysolaminarin is different from callose since it has □-1,6-linked branches. These branches can be cleaved by GH30 CAZymes, which were also detected in low frequencies in the *F*ormosa spp. genomes. Experiments also confirmed *F*ormosa □s ability to grow on laminarin (R. Hahnke, unpublished data). Furthermore, proteomic data confirmed the expression of GH16 enzymes by the *F*ormosa spp. at 07.04.2009.

The genome contents in the sequenced *F*ormosa (draft) genomes also indicate that their ability to synthesize both internal and external □-glucans. The internal □-glucans are also called glycogens and they are a form of carbon reserves for the bacterial cells. The synthesis of internal □-glucans in bacteria is a concerted effort of at least five enzymes: ADP glucose pyrophosphorylase, one glycogen synthase (GT5), one glycogen phosphorylase (GT35), one branching enzyme (GH13), and one debranching enzyme (GH13) [124], and the glucose monomer is delivered by ADP glucose. Glycogen synthases were widely spread among *F*lavobacteriaceae members such as Zobellia galactanivorans Dsij^T [31], Polaribacter MED152 [125] and P. i gensii 23-P [126]. Evidences of glycogen synthesis were found in *F*lavobacterium johnsoniae A3 [127]. Both GH13 and GT5 could be found abundantly in the *F*ormosa □s genomes as well as all *F*ormosa taxobins. But GT35 is absent in the genomes and all its taxobins. The external □-glucans are synthesized by amylosucrases from GH13 and they are usually bundled with CBM48. Amylosucrase takes the glucose moiety from a sucrose molecule and adds it to a □-glucan chain. External □-glucans can be part of the extracellular structures in the *F*ormosa cells. There were four copies of GH13 found in the *F*ormosa taxobin at 07.04.2009. Three of them had CBM48 modules and one of these three carried a signal peptide. The fourth copy was annotated as sucrose synthase. GH13 is only detected in *F*ormosa group A□s genome but not in *F*ormosa group B□s. The external □-glucans were only experimentally observed around *F*ormosa group A cells but not around *F*ormosa group B cells. Together, these results support the hypothesis that *F*ormosa can synthesize □-glucans. However, the absence of GT35

indicated that the incomplete genomes are still not fully studied. It is also possible that *F*ormosa has a novel glycogen metabolic pathway.

A minimum of four enzymes is needed to degrade glycogen: glycogen phosphorylase (GT35), glycogen debranching enzyme (GlgX of GH13), α-1,4-glucanotransferase (amylomaltase GH77), and maltodextrin phosphorylase (GT35) [124]. All these CAZyme families and genes are found in the Reinekea sp. D35 genome and its taxobin at 14.04.2009. In contrast, the Polaribacter draft genomes only contain □-amlyases from GH13. Instead, Polaribacter spp. genomes have higher frequencies of □-glucosidases from GH31 than the Reinekea sp. D35 draft genome. Also, Polaribacter spp. and Reinekea spp. had different transporter profiles. Nevertheless, experiments confirmed that Reinekea sp. D35, Polaribacter sp. 49 and Polaribacter sp. 85 could grow on glycogen. Additionally, Reinekea sp. D35 could grow on mannitol (R. Hahnke, unpublished data). These results suggest that although these two genera employ different biochemical strategies, they both are able to digest external □-glucan. And considering that *F*ormosa could synthesize □-glucan earlier in the bacterioplankton bloom, I speculate that members from both Reinekea and Polaribacter could benefit from this substrate and went into their rapid growth phases. This also explains the delay of their blooms until *F*ormosa built up □-glucans.

Apart from degrading □-glucans, Polaribacter is also equipped with laminarin degradation enzymes. Abundant GH16 and GH5 enzymes were found in Polaribacter spp. genomes and its taxobin at 14.04.2009. In addition, GH30 enzymes were found in all Polaribacter genomes and taxobins. Its laminarin degradation capacity was also proven experimentally (R. Hahnke, unpublished data). However, Polaribacter spp. only possesses low frequencies of sulfatases. Based on these data, I hypothesize that Polaribacter was less able to degrade diatom sulfated extracellular polymers but still capable of digesting the diatom chrysolaminarin. These are also the reasons why Polaribacter did not appear as early as *F*ormosa but had a longer and more intensive bloom than Reinekea in 2009.

It is also worth mentioning that all genomes and taxobins in this study from *F*ormosa and Polaribacter contained both endo- and exo-□-1,3-glucanases. In fact, this is also true for four other marine *F*lavobacteriaceae : *D*okdonia sp. MED134, Leeuwenhoekiella blandensis MED217, *G*ramella forsetii KT0803 [85, 128] and Zobellia galactanivorans Dsij^T [31]. These results suggest that these marine *F*lavobacteriaceae have combined the two synergistic types of enzymes to achieve high laminarin degradation efficiencies.

By putting all observations together, a partial picture of the food web in the spring bloom 2009 became apparent. Diatoms were the primary producers in this event. The

extracellular TEP, the callose strip and the primary storage compound chrysolaminarin in diatoms were all substrates that could be degraded by *F*□□□ □□□ spp.. *F*□□□ □□□ spp. on the other hand synthesized □-glucans both internally and externally. Again, these □-glucans became the substrates for the second wave of bacteria: the □□□□□□□□ and □□□□□□□□□□□. In addition to the □-glucans, □□□□□□□□□□□□spp. were able to benefit from the chrysolaminarin in diatoms.

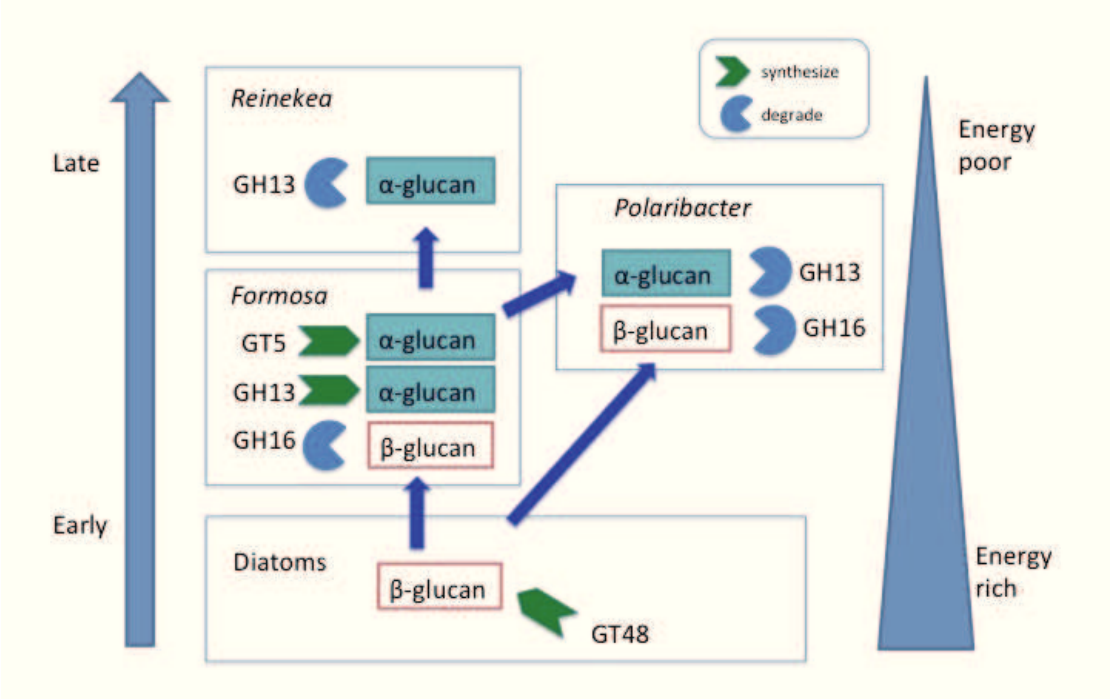


Figure 6. A simplified illustration of the food web in the phytoplankton and bacterioplankton bloom in the North Sea 2009. The arrow on the left indicates the chronicle appearances of the four key players, while the pyramid on the right shows how the energy diminished during the transmission. The central boxes represent the four key players. Inside the boxes are CAZyme families involved in the syntheses and degradations of carbohydrates that linked the adjacent trophic levels. A more detailed description can be found in the main text.

Prior to this thesis, two possible carbohydrate utilization routes were proposed to explain the succession. The first one assumed that different bacterial key players had different substrate spectra. Each one of them benefited from its direct predecessor because the predecessor makes the substrates available. The predecessor either synthesized the substrates or just removed protection layer. As the inorganic or any other resource was depleted, the diatom bloom ended. The primary diatom degraders also decreased simply because of the resulting substrate limitation. They were soon followed by other blooming *B*□□□□□□. However, the organic carbon was converted from labile to recalcitrant states during this process [122, 129] and fewer and fewer biodegradable organic matters were available to sustain the blooms. In contrast, the

second hypothesis stated that the key players were able to degrade similar substrates and the diatom released enough nutrients into water. This model also considered the predecessor as the successor's competitor. When the predecessor was removed from the bacterial population either by predators or by viruses, the successor could rise to fill the vacancy. Therefore, under this hypothesis, mortality via predation and viral attack, not nutrient limitation, were the cause why the bacterial blooms ended. So far, based on the CAZyme data from the key players' taxobins and reference genomes, the succession was better explained by the combination of the two. First, the four key players had a nearly linear carbon and energy transmission. Reinekea spp. lacked laminarinase and no growth on laminarin could be observed experimentally (R. Hahnke, unpublished data). Polaribacter spp. were low in sulfatase frequencies and it might hamper their ability to breach the diatom frustules. These observations were consistent with the first hypothesis' premise. However, both *F*ormosa and Polaribacter were able to utilize laminarin just as the second model expected. To further test these hypotheses, a series of new data is needed, including the densities of predator and viral particles, chrysolaminarin concentrations in the water samples and the β-glucan concentration in the *F*ormosa cells. Also, expression data can reveal the gene product concentrations and help to further clarify the issue.

The relatively simple (if not simplistic) food web structure in the MIMAS metagenome based on CAZyme analyses deserves some attention. Polaribacter spp. were the bacteria that displayed broad polysaccharide degradation capability and could directly interact with the diatom materials. This adds to the complexity of food web and has further implications for this community. Compared with simplistic linear food webs, the complex ones are less prone to exhibit chaotic fluctuation and hence more stable [130]. However, despite Polaribacter spp. could degrade diatom's storage compound, I speculate that *F*ormosa spp. were still indispensible. Since Polaribacter spp. had relatively low frequencies of sulfatases, it was *F*ormosa that cleared the extracellular structures of the diatoms, opened the diatom cells and made the internal chrysolaminarin accessible. In this model, *F*ormosa played the role as a bioconverter from algal polysaccharides to bacterial β-glucans and Reinekea and Polaribacter acted as secondary consumers. And, since β-glucans are easy soluble substrates for a wide variety of organisms, it was possible that the secondary consumers were rather interchangeable. This hypothesis was confirmed by the subsequent observations from 2010 to 2012 at the same sampling site. In these three years, Reinekea spp. were no longer detected to the same extent as in 2009. Instead, other *G*ammaproteobacteria such as Vibrio , *A*lteromonas and *C*olwellia had taken Reinekea's place. CAZyme profiling of their sequenced genomes showed that all these *G*ammaproteobacteria are

enriched in GH13 genes and at the same time possess GH31, GH77 and GT35 genes. These findings fortify the view that the □-glucans from the primary diatom degraders are involved in positive selection for the second wave of bacterioplankton. Any bacteria specialized on degrading □-glucans could replace □□□□□□□□ and bloom on the carbon products from bacteria like *F*□□□ □□□.

This study was the first time that CAZyme profiling was used to investigate a multi-level food web. Different organisms have different compositions of carbohydrates and therefore different sets of CAZymes. Therefore, the CAZyme compositions of organisms within a community could reveal their trophic relations. In practice, the community was captured in the form of metagenome and the organisms within are represented by taxobins. As exemplified here by the MIMAS project, the frequency analyses of the taxobins and reference genomes can connect the dots and generate hypotheses that are experimentally provable. Consequently, the bioinformatic results of this thesis are currently subjected to further experimental testing.

3.4 The different CAZyme profiles in *F*□□□□□□□□□□□□□□□□ reflected their niche adaptations

□□□□□□□□□□□□□□□ and *C*□□□□□□□□□-*F*□□□□□□□□□□□□□-*B*□□□□□□□□□□ are among the most abundant bacteria on Earth [131, 132]. The family *F*□□□□□□□□□□□□□□□□ belongs to the class *F*□□□□□□□□□□□□. □□□□□□□□□, *F*□□□ □□□ and □□□□□□□□□□□, the three genera that appeared in the 2009 North Sea bloom, are members of *F*□□□□□□□□□□□□□□□. As indicated earlier in this thesis or in other studies [133-135], *F*□□□□□□□□□□□ are implicated in the degradation of macromolecules in the ocean.

□□□□□□□□□, *F*□□□ □□□ and □□□□□□□□□□□did not peak simultaneously after the algal bloom in the North Sea in 2009. □□□□□□□□□spp. appeared first and was followed by *F*□□□ □□□ and □□□□□□□□□□ with a one-week delay. This order of succession was observed again in 2010 and 2011 in the North Sea [136]. Hence, I hypothesize that these members of the *F*□□□□□□□□□□□□□□□ have overlapping but distinct enzyme repertoires so that they can adapt to similar but different ecological niches within the course of a diatom bloom. To characterize the ecological niches of these genera, this thesis compared the gene contents and physiologies of these three genera with a broader selection of the *F*□□□□□□□□□□□□□□□ members. In detail, sixteen reference or draft genomes of marine *F*□□□□□□□□□□□□□□□ were searched for CAZymes and sulfatases. They were carried out within the context of algal chemistry, because the life styles of *F*□□□□□□□□□□□□□□□ appear to be in close association with diatoms and brown algae. Four enzyme classes for four algal carbohydrates are chosen for comparison,

namely sulfatases for sulfated polysaccharides, GH16 CAZymes for chrysolaminarin, mannitol dehydrogenases for mannitol and alginate lyases for alginate (Table 3). Experimental observations from R. Hahnke (R. Hahnke, unpublished data) and other studies confirmed the majority of the genomic results (Table 4).

Diatoms and bacteria have coexisted in the oceanic habitats for more than 200 million years. In other words, their interactions, be it synergistic or parasitic, were established over evolutionary time scales [11]. Bacteria were also found closely associated with brown algae, for example bacterial species such as *F* KMM 3553^T have been isolated from them [137]. For this reasons, I speculate that the functional diversification in *F* is potentially linked to the stramenopile evolution. First, sulfated polysaccharides are found in animals and marine organisms, including diatoms and brown algae. Genome comparison indicates that brown algae and animals share some ancestral pathways for the syntheses of sulfated polysaccharides [138]. Their wide distributions in the marine organisms also suggest that polysaccharide sulfatation was an ancient adaptation to the high ionic strength in marine environments [35, 139]. This might explain the wide distributions of sulfatases in *F*, with the exception of *C* HTCC2559^T.

Compared to sulfated polysaccharides, β-1,3-glucans are limitedly distributed. Apart from stramenopiles, they are only found in euglenoids [140, 141], haptophytes [142] as storage compound and in land plants as structural compounds [143]. β-1,3-glucans are found in the form of chrysolaminarin within diatoms and laminarin within brown algae. β-1,3-glucans are degraded by β-1,3-glucanases, which are found in families GH16, GH17 and GH55. Among the sixteen *F* compared, *C.* HTCC2559^T and sp. SCB49 do not contain any sequence from those three families. This is a strong indication that these two bacteria are incapable of degrading β-1,3-glucans.

Alginate and mannitol are found in brown algae but have so far not been observed in diatoms. Alginate, an anionic acid that forms viscous gum when it absorbs water, can make up to 40% of brown algal dry weight. It is found in the extracellular matrix and contributes to the multi-cellularity in brown algae [144]. The mannitol metabolic genes were probably laterally transferred to brown algae from the *A* probably after the separation from diatoms [43, 145]. Among those *F* that do contain GH16 and sulfatases, *D.* Dsij^T, *F.* KMM 3901^T and the two *C* species contain genes for alginate and mannitol recycling, which are absent in sp. 49, *F* group A and B. Other *F* have only one of those two genes, for example *D* MED134 encodes alginate lyases but no mannitol dehydrogenases, while

*K*ordia algicida OT-1 and *C.* atlanticus HTCC2559^{T} was able to oxidize mannitol in growth experiments [146, 147].

Table 3. Sulfatases, mannitol dehydrogenases, GH16 glycoside hydrolases and alginate lyases in sixteen *F*lavobacter aceae . These data have been compiled from this doctoral study and others [56, 85, 88, 125, 126, 128, 146-153]. mannitol dh (dehydrogenase) = Mannitol_dh and Mannitol_dh_C Pfam domains; GH16 = CAZyme family glycoside hydrolase 16; ALs = alginate lyases from CAZyme families PL6, PL7 and PL17. The "+" symbol indicates positive reactions or presence of the genes and "-" indicates the opposite results.

	Sulfatase	mannitol dh	GH16	ALs
*F*ormosa group A	+	-	5	-
*F*ormosa group B	+	-	7	-
F. agari hi a KMM 3901^{T}	+	+	12	+
Po ari acter sp. 49	+	-	6	-
Po ari acter sp. 85	+	+	3	+
P. rgensii 23-P	+	-	4	-
Po ari acter MED152	+	-	5	+
Z. galactanivo ns Dsij^{T}	+	+	15	+
*C*ellulophaga lytica DSM 14237	+	+	8	+
C. al ico a DSM 7489	+	+	8	+
*G*rame la forse ii KT0803	+	+	7	+
U vi act r sp. SCB49	+	-	0	+
C cei acter atlan icus HTCC2559^{T}	-	-	0	+
K a gi ida OT-1	+	-	3	-
*D*okdonia donghaensis MED134	+	-	2	+
Leeuwenhoekiel a bl ndensis MED217	+	+	3	-

Table 4. Sulfated polysaccharide, mannitol, laminarin and alginate utilizations in sixteen *F* tested in growth experiments. These data have been compiled from R. Hahnke (R. Hahnke, unpublished data) and others [56, 85, 88, 125, 126, 128, 146-153]. The '+' symbol indicates positive reactions or presence of the genes and '-' indicates the opposite results. The symbol '?' indicates unknown results.

	mannitol	laminarin	alginate
F group A	-	+	?
F group B	-	+	?
F. KMM 3901^T	+	+	+
sp. 49	-	+	?
sp. 85	-	?	?
. 23-P	-	?	?
MED152	?	?	?
. Dsij^T	+	+	+
C DSM 14237	+	?	?
C. DSM 7489	+	?	?
G KT0803	?	?	?
sp. SCB49	?	?	?
C HTCC2559^T	+	?	-
K OT-1	+	?	-
D MED134	-	?	?
MED217	-	?	?

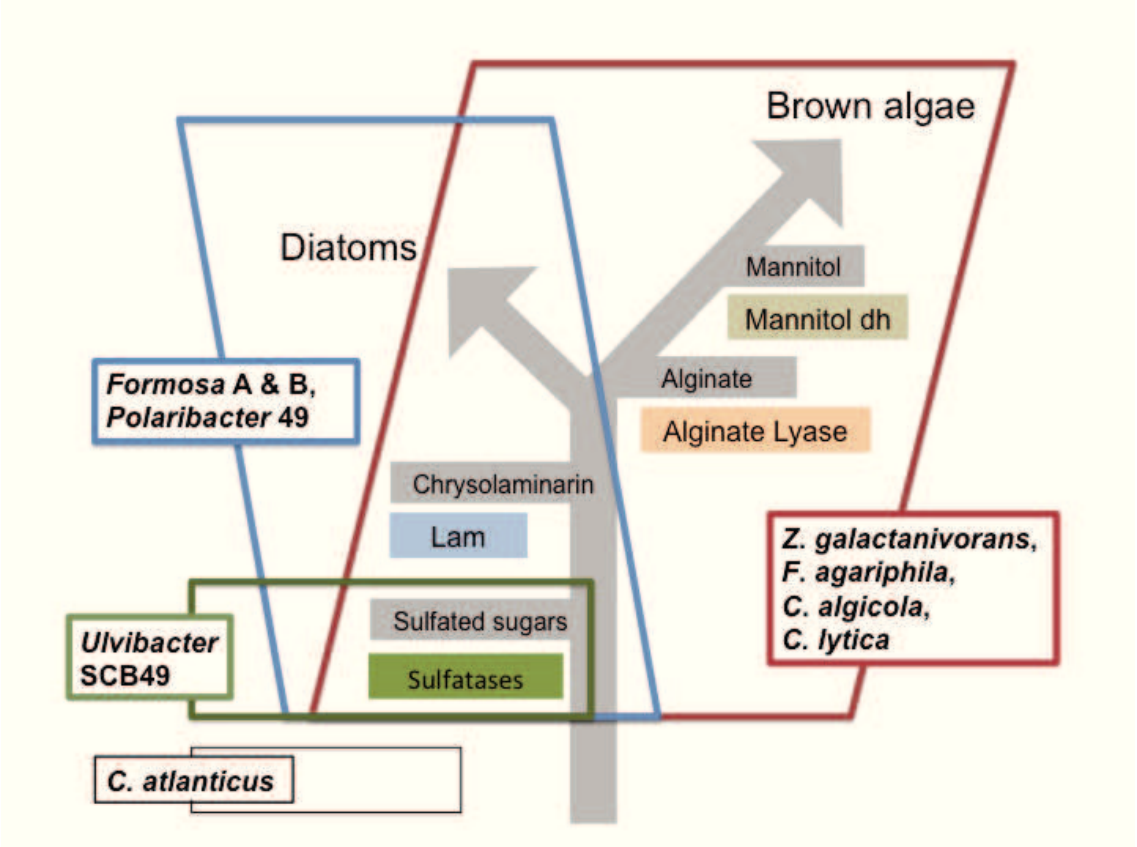


Figure 7. Hypothetical niche differentiation of *F*□□□□□□□□□□□□□□□ and their associations with diatoms and brown algae. This model was based on the carbohydrate compositions of diatoms and brown algae as well as the degradation capacities of sixteen *F*□□□□□□□□□□□□□□□ (Table 3 and Table 4). Four sets of enzymes for the degradation of four algal carbohydrates were chosen: sulfatases for sulfated sugar, laminarinases (Lam) for chrysolaminarin, alginate lyases for alginate and mannitol dehydrogenases (Mannitol dh) for mannitol. The grey parts represent the four algal carbohydrates and their introduction in the evolution of stramenopile. The four colored rectangles represent the enzymes possessed by different *F*□□□□□□□□□□□□□□□ in a Venn diagram manner.

In conclusion, these results suggest that these *F*□□□□□□□□□□□□□□□ occupy different niches (Figure 7). The genome of *C*□□□□□□□□□□□□□□□ HTCC2559^{T} shows the least algal degradation enzymes and I hypothesize that this organism is the least algae-dependent among all the *F*□□□□□□□□□□□□□□□ in this comparison. Its small CAZyme profile did not reveal any specific substrate. □□□□□□□□□ sp. SCB49 also shows few characteristics of algal association. But it contains sulfatases, from which I speculate that it could utilize algal sulfated polysaccharides, especially the easily accessible extracellular ones. This could be one of the causes why □□□□□□□□□ peaked before *F*□□□ □□□ during the spring between 2009 and 2011. The rest of *F*□□□□□□□□□□□□□□□ in this study are all potential diatom degraders, for they all contain GH16 CAZymes that cleave chrysolaminarin (Table 3). However, □. □□□□□□□□□□□□□□□ Dsij^{T}, *F.* □□□□□□□□□ KMM 3901^{T} and the two *C*□□□□□□□□□□ spp. are more likely to be brown algae-

associated. First, these four genomes encode enzymes for the degradations of two carbohydrates found in brown algae but not in diatoms. Second, these four organisms were isolated from algal surfaces [56, 88, 154], whereas □. □□□□□□□23-P, *C.* □□□□□□□□ HTCC2559^T, □□□□□ □□□□□□□□□□□ □□□□□□□□□ MED217, *D*□□□□□□□ □□□□□□□□□□□ MED134 and all North Sea isolates were collected from seawater [126, 147, 152, 153] (R. Hahnke, unpublished data). Finally, it was observed that mannitol partially represses the synthesis of laminarinase in marine □□□□□ [155]. *F*□□□□□□□□□□□□□□□ without mannitol dehydrogenases can suffer from this repression when they interact with the laminarin in brown algae, while the mannitol-converting members can offset this negative effect. In effect, the mannitol in brown algae further differentiates the members in the family *F*□□□□□□□□□□□□□□□ by repelling those without mannitol dehydrogenases. In contrast, *K.* □□□□□□□ OT-1 or *D.* □□□□□□□□□□□ MED134 have partial brown algae degradation capacities and I speculate that they are likely to interact with both diatoms and brown algae.

This niche adaptation model is not completely correlated with the phylogenetic proximity. For example, *F*□□□ □□□ group A and B were more likely to be diatom degraders, while *F.* □□□□□□□□ KMM 3901^T had all the genes to degrade the four chosen carbohydrates in brown algae. Another example was that □□□□□□□□□ sp. 85 is more likely to be associated with brown algae than □□□□□□□□□ sp. 49. Therefore, this model cannot completely reflect the phylogeny of *F*□□□□□□□□□□□□□□□. I hypothesize that the evolution of *F*□□□□□□□□□□□□□□□ is not necessarily parallel to that of the diatoms and brown algae. The present-day genomic landscape of *F*□□□□□□□□□□□□□□□ is imprinted with their associations with different algal lineages. This can be explained by the observation that bacteria can adapt to their habitats fast by limiting their genomes to habitat-specific genes [156]. This study shows that through the careful analyses and interpretations of the CAZymes in *F*□□□□□□□□□□□□□□□, it is possible to pick up these signals left behind during their adaptation processes. Furthermore, it is possible to reversely model the niche separations based on substrate-enzyme relations. However, growth experiments are needed to confirm these hypotheses.

3.5 The CAZyme profiles from the Logatchev metagenome showed that □-glucans and glycosylation are common among the deep-sea microbes.

It was once believed that □-glucans such as starch rarely exist in the deep-sea [157]. However, more recent studies have indicated that □-glucans are also one of the common substrates there [157]. Their metabolic genes are gradually brought to light, especially from those productive hydrothermal vent habitats. During this doctoral

thesis, two hydrothermal vent studies were conducted. The first was a study of shallow-sea hydrothermal vents from the Hot Lake near Panarea, Italy (eighth manuscript; Appendix 7.3). The analysis of Hot Lake samples revealed that thermally stable CAZymes could play a large role in □-glucan recycling. Although GH13 and GT35 could be found abundantly, GH31 were only detected in low frequencies and no GH77 could be identified. Instead, the thermostable □-amylases from GH57 were even more enriched than GH13 amylases. About one fourth of GH57 CAZymes were from *A*□□□□□□ and the rest were bacterial. Most of the enzymes were from extremophilic prokaryotes. Being a hydrothermal vent sample as well, the Logatchev hydrothermal vent field (LHF) metagenome might share some common characteristics with Hot Lake. For this reason, the studies of Hot Lake have provided important hints about the life at hydrothermal vents.

About 1% of the genes in LHF metagenome were encoding CAZymes. Unlike Hot Lake metagenomes, LHF metagenome had a higher GH13:GH57 ratio (10:3). This was perhaps because the deep-sea LHF had a low temperature (4□C) than Hot Lake (36□C and 74□C at two different sites). However, LHF still had a GH57 frequency (0.01%) that was higher than the metagenomes from the MIMAS study. This intermediate GH57 frequency could indicate that the LHF community was a mixture of the hot-adapted and cold-adapted microbes.

The second finding in the LHF CAZymes was that protein glycosylation is also common in the deep-sea, which had never been reported before. Several glycosylation CAZyme families were identified. Examples included GH109, GT27, GT39, GT41 and GT66. GT27 enzymes initiate mucin-type O-glycosylation in animals [158], but they were also identified in prokaryotes. GT39 contains dolichyl-phosphate-mannose-protein mannosyltransferases that are responsible for the O-linked glycosylation of proteins. The presences of these CAZyme families hinted that deep-sea prokaryotes are also capable of protein glycosylation. Various studies showed that protein glycosylation in bacteria could modify the underlying proteins in different ways, such as to increase the enzyme heat stability and affinity.

The presences of □-glucan metabolism and protein glycosylation in the LHF metagenome showed that the deep-sea microbial community is not alien in comparison to surface water counterparts. However, more research is needed to provide new insight into the microbial activities at these deep-sea habitats.

3.6 CAZyme profiling as tool for the study of microbial interactions in meta-omics

Field studies are sometimes much harder to conduct than laboratory experiments. Because the conditions in the field can hardly be controlled, careful design is needed so that the confounding factors can be accounted for. Furthermore, some of the powerful laboratory techniques prove to be less useful on a large scale. The tools in a field microbiologist's toolkit are rather limited. Meta-omics studies provide snapshots of the biotic elements in a habitat. Complemented by physicochemical measurement, they can provide a rather holistic view on a microbial ecosystem. In an ecosystem, organisms are not isolated. They interact with each other as well as with their surrounding environment. These abilities for interactions are programmed in their genomes and implemented by their expressed gene products, such as functional RNAs and proteins. Meta-omics studies can potentially capture all these genomic potentials, expression patterns and the interactions. However, the raw data from a meta-omic study are nothing more than just some sequences and signals. To distill the interactions, especially the trophic connections, a higher level of understanding of these raw data is needed. Taxonomic classification and functional annotation of sequence data lay the foundation for such an understanding. However, among thousands of functional proteins, the majority of which is involved in cell maintenance and only part of that repertoire is involved in cell-cell trophic interactions. A small portion of the genome is dedicated to the carbohydrate metabolism. These CAZymes are one of the well-characterized genes. Their substrates, the carbohydrates, are ubiquitous and involved in the major biochemical pathways in life. As more and more studies have already indicated, carbohydrates are an essential part of the molecule exchange network in a microbial community. And CAZyme profiling is the tool to chart it. Without doubt, the characterization of CAZyme will remain one of the most rewarding and revealing analyses in meta-omic studies.

4. O□□□□□□

Although this thesis describes the trophic connections among four key players in North Sea in spring 2009, they are just a small part of a far bigger and much more complex food web. Several other bacterial genera were also found abundantly. They include □□□□□□□□□ spp., members of the class *F*□□□□□□□□□□ that bloomed between the chlorophyll a peak and the *F*□□□ □□□ bloom. Also, the alphaproteobacterial □□□□□□□□□□and SAR11 clades accounted for substantial proportions of the bacterial population. The two gammaproteobacterial clades SAR86 and SAR92, bloomed about three weeks after the □□□□□□□ bloom [15, 136]. Although it was beyond the scope of this doctoral thesis, the relations among these organisms have yet to be explored. It is also important to realize that exploratory (meta-)genomics and CAZyme profiling results □only□in hypotheses, which need further examination. It is for example possible to test the □-glucan production of various *F*□□□ □□□ strains via a high-throughput screening platform based on microarrays populated with well-defined poly- and oligosaccharides [159]. To prove the interdependency of □□□□□□□ and □□□□□□□□□on *F*□□□ □□□s □-glucans, crossfeeding experiments in enrichments or defined mixed cultures can be conducted. The results can then be visualized by FISH.

The method described in this thesis □ Trident, CAZyme frequency analysis, substrate availability study and bioinformatics-guided experiments □ represents a first attempt to analyze CAZymes in a large genomic dataset. The complete workflow has been revised and improved along with the MIMAS Project and the Logatchev study. However, the pipeline still has room for improvement. The following two ideas should be considered to be included:

First, neither dbCAN nor BLAST places any emphasis for the catalytic sites of the CAZyme hits. The Prosite website has currently listed 41 signature patterns for 30CAZyme families. Most of those patterns capture the catalytic sites of the families and they are vital for the automatic classifications. This information can be incorporated into the pipeline to eliminate some false positive results and to highlight the true positive results. The Prosite signatures can be coded into Trident as part of the high-throughput automatic screening pipeline.

Second, the Phyre2 website offers a good □□ □□□□□ 3D fold prediction of proteins. Although Phyre2 by no means can replace X-ray crystallography or NMR for structural determination, it does add another layer of insight into enzyme. Besides, it is free, easy and relatively fast. It is even recommended to use Phyre2 to validate enzymes before they are prepared for laboratory experiments. However fast, Phyre2 is rather

computationally intensive and it typically takes more than three hours to complete the analysis of one protein. For this exact reason, the server sets limit to the submitted job to avoid overload. Phyre2, however, does provide a batch processing capacity for academic users. It is thus possible to automatically compile a list of good candidate CAZymes after Trident analysis and submit them for batch processing.

The automatic identification of CAZymes by Trident is considered to be the first step in metagenomic studies. It is also an informative process that quickly outlines the characteristics of the underlying metagenome. Frequency analyses can point out abundant or rare CAZymes that deserve further attention. Any in-depth study should not stop on the level of frequency results. The substrates, the related pathways and the habitats of the community should be also considered to get a more holistic view of the carbohydrate metabolism.

It is also possible to predict the substrate specificities of CAZymes. For that it is assumed that enzymes with similar sequences also have similar substrates. The similarities of sequences can be deduced either from the bit scores when they are compared through BLAST or from the topologies in phylogenetic analyses. Bit scores can be directly read from the BLAST results. Phylogenetic analysis, on the other side, requires more dedicated effort. First, candidate sequences are classified into CAZyme families. Afterwards, reference sequences of the families with known substrate specificities on the CAZy website are collected. Then phylogenetic trees can be reconstructed from alignments. The tree algorithm is a matter of choice. It depends on factors such as the amount of sequences, time and resource constraint. The phylogenetic trees can be visually examined by tools such as iTOL [160] and the neighbors of candidate sequences can be determined. Finally, the interpretation of CAZyme trees is not different from a 16S phylogenetic tree. The substrate specificities can be deduced based on those of the neighbors. Several studies provided good examples of CAZyme phylogenetic analyses [51, 55, 82].

Aside from these ecological follow-up studies, Trident can also be used as a biotechnological data-mining tool for industrial applications. CAZymes attract more and more interest as biotechnology develops rapidly. CAZyme shows great promise in the next round of technological breakthroughs in fields such as agriculture, the manufacturing of products like cosmetics and paper, second generation biofuels, waste treatment and medicine. Before any upscale applications, candidate CAZymes should be examined in details both □□ □□□□□□ and then □□ □□□□□. This computational examination is the same as in all the studies described in this thesis, except the selection is more stringent. In fact, in the genome project of archaeon *H*□□□□□□□□□□ □□□□ □□□□ SARL4B^{T}, the pipeline has already been applied for data mining of CAZymes with

biotechnological potential. Among all identified CAZymes, four were selected. There are two GH5, one GH43 and one GH13. All these genes were translated and analyzed by Phyre2. The protein sequences were aligned with characterized CAZyme sequences and afterwards their substrate specificities were determined by phylogenetic reconstructions. These four enzymes are currently prepared for cloning. Afterwards, the □□□□□□ results can be funneled to a medium throughput cloning and expression study, which produces soluble candidate proteins for the subsequent functional and structural characterization in experiment [161, 162]. The results will greatly advance our understanding of the biochemical properties of the enzymes and prepare us for the next-stage up-scale production.

Since the advent of CAZy database, CAZyme studies have evolved from simple profiling to the more elaborate and multi-disciplinary research projects involving crystallography, informatics, physic, chemistry and biology. These sophisticated projects have the potential to much more accurately characterize CAZymes. Their results open up more and more opportunities for industrial application. The proposed improvements in this chapter are small steps towards a modern analytical pipeline to provide high-quality CAZymes profiles. Hopefully, the refined Trident will eventually becomes a valuable tool in the metagenomic toolkit.

5. Refe□e□ce□ ci□ed i□ l□□□□d□c i□□, Di□c□□□i□□ a□d O□□□□□□

1. **He□i□□a□ B, C□□i□h□ PM, Da□ie□ GJ:** A census of carbohydrate-active enzymes in the genome of *A□□□□□□□□□□□□ □□□□□□□□□*. □□□□□□ □□*B*□□□ 2001, **47**:55-72.
2. **C□□i□h□ PM, De□e□□□ E, Da□ie□ GJ, He□i□□a□ B:** An evolving hierarchical family classification for glycosyltransferases. *J* □ □□*B*□□□ 2003, **328**:307-317.
3. **C□□i□h□ PM, S□a□ M, B□a□c E, He□i□□a□ B:** Why are there so many carbohydrate-active enzyme-related genes in plants? □□□□□□ □□□□□ □□□ 2003, **8**:563-565.
4. **Lai□□□□ LL, He□i□□a□ B, Da□ie□ GJ, Wi□he□□ SG:** Glycosyltransferases: structures, functions, and mechanisms. *A*□□□ □ □□*B*□□□□□□ 2008, **77**:521-555.
5. **Cha□e□□e MA, S□i□h WH:** The volume of Earth's ocean. □ □□□□□□□□□□□□ 2010, **23**:112-114.
6. **M□□a C, Ti□□e□□□□ DP, Ad□S, Si□ □□□□ AG, W□□□ B:** How many species are there on Earth and in the ocean? □□□□ □□□□□□ 2011, **9**:e1001127.
7. **Ma□□ DG:** The species concept in diatoms. □□□□□□□□□ 1999, **38**:437-495.
8. **Fie□d CB, Beh□e□fe□d MJ, Ra□de□□□□ JT, Fa□□□□ □□i P:** Primary production of the biosphere: integrating terrestrial and oceanic components. □□□□□□□ 1998, **281**:237-240.
9. **Fa□□□□ □□i PG, Ba□be□ RT, S□e□ace□ V:** Biogeochemical Controls and Feedbacks on Ocean Primary Production. □□□□□□□ 1998, **281**:200-207.
10. **K□□h PG, Chi□□i□i A, G□□be□ A, Ma□i□-Je□e□□e□V, M□c□ T, Pa□□e□ MS, S□a□□e□ MS, Ka□□a□ A, Ca□□□ L, Webe□ T, Mahe□□a□i U, A□□b□□□□ EV, B□□□e□ C:** A model for carbohydrate metabolism in the diatom □□□□□□□□□□□□□□ deduced from comparative whole genome analysis. □□□□ □ □□ 2008, **3**:e1426.
11. **A□i□ SA, Pa□□e□MS, A□□b□□□□EV:** Interactions between diatoms and bacteria. □ □□□□□□□□□ □□*B*□□□ □□ 2012, **76**:667-684.
12. **Lig□e□□ R, Hei□□a□e□ AS, K□□□a H, G□□de□□e□ K, K□□□□□-Lei□i□□i P, Paj□□ie□i R, Ui□□□ A:** Fate of a phytoplankton spring bloom: sedimentation and carbon flow in the planktonic food web in the northern Baltic. □ □□□□ *E*□□□□□□-□□□□□□□□ □ □□□□ 1993, **94**:239.

13. **Shimada A, Shiga N, Ban S:** Origin of [illegible] diatoms that cause the spring phytoplankton bloom in Funka Bay, southwestern Hokkaido, Japan. [illegible] 1999, **46**:89-93.
14. **Maier G, Glegg GA, Tappin AD, Worsfold PJ:** A high resolution temporal study of phytoplankton bloom dynamics in the eutrophic Taw Estuary (SW England). [illegible] 2012, **434**:228 - 239.
15. **Teeling H, Fuchs BM, Becher D, Klockow C, Gardebrecht A, Bennke CM, Kassabgy M, Huang S, Mann AJ, Waldmann J, Weber M, Klindworth A, Otto A, Lange J, Bernhardt J, Reinsch C, Hecker M, Peplies J, Bockelmann FD, Callies U, Gerdts G, Wichels A, Wiltshire KH, Glöckner FO, Schweder T, Amann R:** Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. [illegible] 2012, **336**:608-611.
16. **Jannasch HW, Wirsen CO, Nelson DC, Robertson LA:** [illegible] sp. nov., a colorless, sulfur-oxidizing bacterium from a deep-sea hydrothermal vent. [illegible] 1985, **35**:422-424.
17. **Campbell BJ, Jeanthon C, Kostka JE, Luther GW, Cary SC:** Growth and phylogenetic properties of novel bacteria belonging to the epsilon subdivision of the [illegible] enriched from [illegible] and deep-sea hydrothermal vents. [illegible] 2001, **67**:4566-4572.
18. **Campbell BJ, Engel AS, Porter ML, Takai K:** The versatile epsilon-proteobacteria: key players in sulphidic habitats. [illegible] 2006, **4**:458-468.
19. **Azam F, Fenchel T, Field JG, Gray JS, Meyer-Reil LA, Thingstad F:** The Ecological Role of Water-Column Microbes in the Sea. [illegible] 1983, **10**:257-263.
20. **Azam F:** Microbial Control of Oceanic Carbon Flux: The Plot Thickens. [illegible] 1998, **280**:694-696.
21. **Azam F, Malfatti F:** Microbial structuring of marine ecosystems. [illegible] 2007, **5**:782-791.
22. **Hollibaugh JT, Azam F:** Microbial degradation of dissolved proteins in seawater. [illegible] 1983, **28**:1104-1116.
23. **Verdugo P, Alldredge AL, Azam F, Kirchman DL, Passow U, Santschi PH:** The oceanic gel phase: a bridge in the DOM--POM continuum. [illegible] 2004, **92**:67-85.
24. **Verdugo P:** Marine microgels. [illegible] 2012, **4**:375-400.
25. **D[illegible]ch DS:** Algal Cell Walls. In [illegible]. John Wiley & Sons, Ltd; 2001.

26. **S[illegible]ic[illegible] RV, a[illegible]d Wi[illegible]ia[illegible] SJ:** *C[illegible]*. Amsterdam; London: Elsevier; 2009.

27. **Lai[illegible]e RA:** A calculation of all possible oligosaccharide isomers both branched and linear yields 1.05 x 10(12) structures for a reducing hexasaccharide: the Isomer Barrier to development of single-method saccharide sequencing or synthesis systems. *G[illegible]* 1994, **4**:759-767.

28. **Be[illegible]ea[illegible] O, M[illegible] B:** Sulfated fucans, fresh perspectives: structures, functions, and biological properties of sulfated fucans and an overview of enzymes active toward this class of polysaccharide. *G[illegible]* 2003, **13**:29R-40R.

29. **Ba[illegible]be[illegible] T, Ge[illegible]a[illegible]d A, P[illegible]i[illegible] P, He[illegible]i[illegible]a[illegible] B, K[illegible]a[illegible]eg B:** The kappa-carrageenase of the marine bacterium *C[illegible]* drobachiensis. Structural and phylogenetic relationships within family-16 glycoside hydrolases. *[illegible] B[illegible] E[illegible]* 1998, **15**:528-537.

30. **Miche[illegible]G, N[illegible]a[illegible]-C[illegible]e[illegible] P, Ba[illegible]be[illegible] T, C[illegible]j[illegible]e[illegible] M, He[illegible]be[illegible] W:** Bioconversion of red seaweed galactans: a focus on bacterial agarases and carrageenases. *A[illegible] B[illegible]* 2006, **71**:23-33.

31. **Hehe[illegible]a[illegible] J-H, C[illegible]ec G, Ba[illegible]be[illegible] T, He[illegible]be[illegible] W, C[illegible]j[illegible]e[illegible] M, Miche[illegible]G:** Transfer of carbohydrate-active enzymes from marine bacteria to Japanese gut microbiota. *[illegible]* 2010, **464**:908-912.

32. **Lee JB, Ha[illegible]a[illegible]hi K, Hi[illegible]a[illegible]a M, K[illegible]da E, S[illegible]i E, K[illegible]b[illegible] Y, Ha[illegible]a[illegible]hi T:** Antiviral sulfated polysaccharide from *[illegible]*, a diatom collected from deep-sea water in Toyama Bay. *B[illegible] B[illegible]* 2006, **29**:2135-2139.

33. **Laha[illegible]e M, R[illegible]bic A:** Structure and functional properties of ulvan, a polysaccharide from green seaweeds. *B[illegible]* 2007, **8**:1765-1774.

34. **Miche[illegible] G, T[illegible] T, Sc[illegible]e[illegible] D, C[illegible]c[illegible] JM, K[illegible]a[illegible]eg B:** The cell wall polysaccharide metabolism of the brown alga *E[illegible]*. Insights into the evolution of extracellular matrix polysaccharides in Eukaryotes. *[illegible]* 2010, **188**:82-97.

35. **A[illegible]i[illegible] RS, G[illegible]a[illegible]i[illegible]C, M[illegible] PAS:** Rising from the sea: correlations between sulfated polysaccharides and salinity in plants. *[illegible]* 2011, **6**:e18862.

36. **Da[illegible]a[illegible]-Sa[illegible] N, G[illegible]e[illegible] DL, C[illegible]a LS, C[illegible]dei[illegible] SL, C[illegible]a MS, T[illegible]i[illegible]dade ES, F[illegible]a[illegible]c[illegible] CR, Sc[illegible]ecci KC, Lei[illegible]e EL, R[illegible]cha HA:** Freshwater Plants Synthesize Sulfated Polysaccharides: Heterogalactans from Water Hyacinth (*E[illegible]*). *I[illegible] J [illegible]* 2012, **13**:961-976.

37. **K[illegible]aa[illegible] S:** Algal Polysaccharides, Novel Applications and Outlook. In *C[illegible] - C[illegible] G[illegible] G[illegible]*.

1st edition. Edited by Chang C. Janeza Trdine 9, 51000 Rijeka, Croatia: InTech; 2012:489-532.

38. **A□d□edge AL:** The potential role of particulate diatom exudates in forming nuisance mucilaginous scums. *A□□ I□□□□□□□□□□□* 1999, **35**:397-400.

39. **Bea□ie A, Hi□□□ EL, Pe□ci□a□ E:** Studies on the metabolism of the *C□□□□□□□□□□e*. Comparative structural investigations on leucosin (chrysolaminarin) separated from diatoms and laminarin from the brown algae. *B□□□□□ J* 1961, **79**:531-537.

40. **Ha□da N:** Carbohydrate metabolism in the marine diatom □□□□□□□□□ □□□□□□□. □ □□□□ *B□□□□* 1969, **4**:208-214.

41. **V□□□□ KM, □□gaa□d K, G□i□□□□d K:** Diurnal rhythms in carbohydrate metabolism of the marine diatom □□□□□□□□□ □ □□□□□□□ (Grev.) Cleve. *J□□□□□* □□□□□□□□ □□□□□□ □□□□ □□□□□□ □□□ □□□□□□ 1986, **102**:249-256.

42. **M□□e□ad S, E□□e□ G:** Biology of (1,3)-□-glucans and related glucans in protozoans and chromistans. In *C□□□ □□□, B□□□□□ □□□, □□□ B□□□□□ □□1-3 B□□ G□□□□□ □□□ □□□□□□ □□□□□□□□□□□□□.* 1st edition. Edited by Bacic A, Fincher G, Stone B. Academic Press/Elsevier; 2009:353-385.

43. **Di□a□i SM, Aa□ HT, Pa□□e□ BS, B□□e□ C, Ed□a□d□e□ B, T□□□□ T:** Mannitol in six autotrophic stramenopiles and Micromonas. □□□□ □□□□□ *B□□□□* 2011, **6**:1237-1239.

44. **A□de□a□ □ A:** Carbohydrate production by phytoplankton and degradation in the marine microbial food web. □□*D* □□□□. Rijksuniversiteit Groningen; 2006.

45. Callose dimer [http://commons.wikimedia.org/wiki/File:Callose.svg]

46. **Ca□□a□e□BL, C□□□i□h□ PM, Ra□c□□e□C, Be□□a□d T, L□□ba□d V, He□□i□□a□B:** The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. □□□□□ *A□□□ □□□* 2009, **37**:D233-D238.

47. **Da□ P, Ka□ae□a I, Ya□g S-J, Zh□□ F, Yi□ Y, Ch□□ W, P□□□e FL, We□□he□i□g J, He□ich R, Gia□□□□e R, Le□i□ DL, Ke□□ R, Gi□be□□HJ, He□□i□□a□B, X□ Y, Ada□□ MWW:** Insights into plant biomass conversion from the genome of the anaerobic thermophilic bacterium *C□□□□□□□□□□□□□□□ □□□□□* DSM 6725. □□□□□ *A□□□ □□□* 2011, **39**:3240-3254.

48. **B□a□□□□ AB, B□a□ DN, Gi□be□□ HJ, Da□ie□ GJ:** Carbohydrate-binding modules: fine-tuning polysaccharide recognition. *B□□□□□ J* 2004, **382**:769-781.

49. **Da□ie□ G, He□□i□□a□B:** Structures and mechanisms of glycosyl hydrolases. □□□□□□□ 1995, **3**:853 - 859.

50. **He□□i□□a□ B, Ca□eba□□ I, Fab□ega S, Leh□ P, M□□□□□ J-P, Da□ie□ G:** Conserved catalytic machinery and the prediction of a common fold for several

families of glycosyl hydrolases. *Proceedings of the National Academy of Sciences* 1995, **92**:7090-7094.

51. **Stam MR, Danchin EGJ, Rancurel C, Coutinho PM, Henrissat B:** Dividing the large glycoside hydrolase family 13 into subfamilies: towards improved functional annotations of alpha-amylase-related proteins. *Protein Eng Des Sel* 2006, **19**:555-562.
52. **Xu J, Bjursell MK, Himrod J, Deng S, Carmichael LK, Chiang HC, Hooper LV, Gordon JI:** A genomic view of the human-Bacteroides thetaiotaomicron symbiosis. *Science* 2003, **299**:2074-2076.
53. **Mahowald MA, Rey FE, Seedorf H, Turnbaugh PJ, Fulton RS, Wollam A, Shah N, Wang C, Magrini V, Wilson RK, Cantarel BL, Coutinho PM, Henrissat B, Crock LW, Russell A, Verberkmoes NC, Hettich RL, Gordon JI:** Characterizing a model human gut microbiota composed of members of its two dominant bacterial phyla. *Proc Natl Acad Sci U S A* 2009, **106**:5859-5864.
54. **Chassard C, Delmas E, Lawson PA, Bernalier-Donadille A:** *Bacteroides xylanisolvens* sp. nov., a xylan-degrading bacterium isolated from human faeces. *Int J Syst Evol Microbiol* 2008, **58**:1008-1013.
55. **Strohmeier M, Hrmova M, Fischer M, Harvey AJ, Fincher GB, Pleiss J:** Molecular modeling of family GH16 glycoside hydrolases: potential roles for xyloglucan transglucosylases/hydrolases in cell wall modification in the poaceae. *Protein Sci* 2004, **13**:3200-3213.
56. **Barbeyron T, L'Haridon S, Corre E, Kloareg B, Potin P:** *Zobellia galactanovorans* gen. nov., sp. nov., a marine species of *Flavobacteriaceae* isolated from a red alga, and classification of [*Cytophaga*] uliginosa (ZoBell and Upham 1944) Reichenbach 1989 as *Zobellia uliginosa* gen. nov., comb. nov. *Int J Syst Evol Microbiol* 2001, **51**:985-997.
57. **Martens EC, Lowe EC, Chiang H, Pudlo NA, Wu M, McNulty NP, Abbott DW, Henrissat B, Gilbert HJ, Bolam DN, Gordon JI:** Recognition and degradation of plant cell wall polysaccharides by two human gut symbionts. *PLoS Biol* 2011, **9**:e1001221.
58. **Kostylev M, Wilson D:** Synergistic interactions in cellulose hydrolysis. *Biofuels* 2012, **3**:61-70.
59. **Henrissat B, Driguez H, Viet C, Schülein M:** Synergism of cellulases from *Trichoderma reesei* in the degradation of cellulose. *Nat Biotechnol* 1985, **3**:722-726.
60. **Coughlan MP, Moloney AP, McCrae SI, Wood TM:** Cross-synergistic interactions between components of the cellulase systems of *Talaromyces emersonii*

emersonii , *F*usarium solani , Penicillium funiculosum , and Trichoderma koningii *B*iochem . Soc . Trans 1987, **15**:263-264.

61. **Ja□a□ J, K□□a□i□ M, Te□gja□ H, V□ja□ □e P:** Endo-exo synergism in cellulose hydrolysis revisited. *J B*iol *C*hem 2012, **287**:28802-28815.
62. **G**us**a**kov **AV, Si**n**i**tsyn **AP:** A theoretical analysis of cellulase product inhibition: effect of cellulase binding constant, enzyme/substrate ratio, and beta-glucosidase activity on the inhibition pattern. *B*iotechnol *B*ioeng 1992, **40**:663-671.
63. **D**risk**i**ll **LE, Ba□e□ MW, Ke**lly **RM:** Synergistic interactions among beta-laminarinase, beta-1,4-glucanase, and beta-glucosidase from the hyperthermophilic archaeon Pyrococcus furiosus during hydrolysis of beta-1,4-, beta-1,3-, and mixed-linked polysaccharides. *B*iotechnol *B*ioeng 1999, **66**:51-60.
64. **Laf□□d M, Na□a□□□ D, Ha□□ M, C**outuri**e**r **M, Be**rrin **JG:** Characterization of a broad-specificity □-glucanase acting on β-(1,3)-, β-(1,4)-, and □-(1,6)-glucans that defines a new glycoside hydrolase family. *A*ppl *E*nviron Microbiol 2012, **78**:8540-8546.
65. **F□□□da K, Hi□aga M, A□a□□□a S, A□ai I, Se□i□a□a M, U□a□hi□a T:** Purification and characterization of a novel exo-beta-1,3-1,6-glucanase from the fruiting body of the edible mushroom Enoki (*F*lammulina velutipes). *B*iosci *B*iotechnol *B*iochem 2008, **72**:3107-3113.
66. **Need□e□a□ SB, W□□□ch CD:** A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J* Mol *B*iol 1970, **48**:443-453.
67. **S□i□h TF, Wa□e□□a□ MS:** Identification of common molecular subsequences. *J* Mol *B*iol 1981, **147**:195-197.
68. **A**lts**ch**ul **SF, Gi□h W, Mi□e□W, M**y**e**rs **EW, Li□□a□ DJ:** Basic local alignment search tool. *J* Mol *B*iol 1990, **215**:403-410.
69. **K□□gh A, B**rown **M, Mia□ IS, Sj□□a□de□K, Ha□□□□e□D:** Hidden Markov models in computational biology. Applications to protein modeling. *J* Mol *B*iol 1994, **235**:1501-1531.
70. **Edd□ SR:** A probabilistic model of local sequence alignment that simplifies statistical significance estimation. PLoS *C*omput *B*iol 2008, **4**:e1000069.
71. **Edd□ SR:** Profile hidden Markov models. *B*ioinformatics 1998, **14**:755-763.
72. **K□a□□e L, Dia□ NN, G**oe**sm**ann **A, Ke□□e□ S, Na**ttk**e**mp**e**r **TW, R□h□e□ F, Ed□a□d□ RA, S**toy**e J:** Phylogenetic classification of short environmental DNA fragments. Nucleic *A*cids Res 2008, **36**:2230-2239.
73. **S**onn**ha**mm**e**r **EL, Edd□ SR, D□□bi□ R:** Pfam: a comprehensive database of protein domain families based on seed alignments. Proteins 1997, **28**:405-420.

74. **S□□□ha□ □e□EL, Edd□ SR, Bi□□e□ E, Ba□e□ a□ A, D□□bi□ R:** Pfam: multiple sequence alignments and HMM-profiles of protein domains. □□□□□□□ *A□□□□* □□□ 1998, **26**:320-322.

75. **Fi□□ RD, Ta□e J, Mi□□□□ J, C□ggi□ PC, Sa□ □ □□ SJ, H□□□ HR, Ce□ic G, F□□□□□□d K, Edd□ SR, S□□□ha□ □e□ EL, Ba□e□ a□ A:** The Pfam protein families database. □□□□□□□ *A□□□□* □□□ 2008, **36**:D281-D288.

76. **Fi□□ RD, Mi□□□□ J, Ta□e J, C□ggi□ P, Hege□ A, P□□i□g□□□ JE, Ga□i□ OL, G□□a□e□a□a□ P, Ce□ic G, F□□□□□□d K, H□□□ L, S□□□ha□ □e□ELL, Edd□ SR, Ba□e□ a□ A:** The Pfam protein families database. □□□□□□□ *A□□□□* □□□□□□□□ 2010, **38**:D211-D222.

77. **Gei□e□-Lee J, Gei□e□ M, C□□i□h□ PM, Sege□□a□ B, Ni□hi□□b□ N, Ta□aha□hi J, A□□eb□g H, Dje□bi S, Ma□e□ E, A□de□□□□□-G□□□e□bac□□a□h□a□ S, S□□dbe□g B, Ka□□i□□□i S, Tee□i TT, K□ec□□□□ □□i LA, He□i□□a□ B, Me□e□□□□ic□ EJ:** Poplar carbohydrate-active enzymes. Gene identification and expression analyses. □□□□□□□ □□□□□□ 2006, **140**:946-962.

78. **Ta□□e L, Be□c□□ici J, Pi□□□□-Se□i□ S, R□be P, Ta□ J, K□□□□ C, Ca□□a□e□BL, C□□i□h□ PM, He□i□□a□B, Lec□e□c M, D□□□ J, M□□□a□ P, Re□a□d-Si□e□□ M, P□□□c□i-Ve□□□e□e G:** Functional metagenomics to mine the human gut microbiome for dietary fiber catabolic enzymes. *G□□□□* □ □□□ 2010, **20**:1605-1612.

79. **He□□ M, Sc□□□ba A, Ega□ R, Ki□ T-W, Ch□□ha□a□a H, Sch□□□h G, L□□ S, C□a□□ DS, Che□ F, Zha□g T, Mac□ie RI, Pe□□acchi□ LA, T□i□ge SG, Vi□e□A, W□□□e T, Wa□g Z, R□bi□ EM:** Metagenomic discovery of biomass-degrading genes and genomes from cow rumen. □□□□□□□ 2011, **331**:463-467.

80. **Pa□□ BH, Ka□□i□e□□ TV, S□ed MH, Le□□e MR, Ube□bache□ EC:** CAZymes Analysis Toolkit (CAT): web service for searching and analyzing carbohydrate-active enzymes in a newly sequenced organism using CAZy database. *G□□□□□□□□□□□* 2010, **20**:1574-1584.

81. **Yi□ Y, Ma□ X, Ya□g J, Che□ X, Ma□ F, X□ Y:** dbCAN: a web resource for automated carbohydrate-active enzyme annotation. □□□□□□□ *A□□□□* □□□ 2012, **40**:W445-W451.

82. **Che□ W, Xie T, Sha□ Y, Che□ F:** Phylogenomic Relationships between Amylolytic Enzymes from 85 Strains of Fungi. □□□□ □ □□ 2012, **7**:e49679.

83. **Sig□□□CJ, de Ca□□□□ E, Ce□□□□i L, C□che BA, H□□□ N, B□idge A, B□□g□e□e□e□ L, Xe□a□i□□ I:** New and continuing developments at PROSITE. □□□□□□□ *A□□□□* □□□ 2013, **41**:D344-D347.

84. **Ke□e□ LA, S□e□□be□g MJE:** Protein structure prediction on the Web: a case study using the Phyre server. □□□□ □□□□□ 2009, **4**:363-371.

85. **Ba□e□M, K□be M, Tee□i□g H, Rich□e□M, L□□ ba□d□□T, A□e□□ E, W□□de□ a□□ CA, Q□a□□ C, K□h□H, K□a□□□ F, W□eb□e□ D, Bi□ch□f K, M□□□□ a□□ M, Ch□□dh□□i JV, Me□e□ F, Rei□ha□d□ R, A□a□□ Rl, G□□c□□e□ FO:** Whole genome analysis of the marine Bacteroidetes '*G□□□ □□□ □□□□□□□*' reveals adaptations to degradation of polymeric organic matter. *E□□□□□ □ □□□□□□□□* 2006, **8**:2201-2213.

86. **Ab□B, L□ M, Mi□a M, Ha□ C, N□a□ M, L□ca□ S, Ha□ □ □□ N, De□h□a□de S, Che□g J-F, Ta□ia R, G□□d□i□ L, Pi□□□c□ S, Li□i□□ K, Paga□i l, l□a□□□a N, Ma□□□□ a□i□ K, O□chi□i□□□a G, Pa□i A, Che□ A, Pa□a□ia□□a□ K, La□d M, Ha□□e□L, Cha□g Y-J, Jeff□ie□ CD, De□□e□JC, B□a□ bi□a E, R□hde M, Ti□da□□ BJ, G□□e□M, W□□□e T, B□i□□□□ J, Ei□e□ JA, Ma□□□□i□□ V, H□ge□h□□□ P, K□□□ide□ NC, K□e□□ H-PP-PP, La□id□□ A:** Complete genome sequence of *C□□□□□□□□□□□ □□□□□□□□* type strain (IC166). *□□□□□ G□□□□ □□ □□□* 2011, **4**:72-80.

87. **Pa□i A, Ab□ B, Te□hi□a H, N□a□ M, La□id□□ A, L□ca□ S, Ha□ □ □□ N, De□h□a□de S, Che□g J-F, Ta□ia R, Ha□ C, G□□d□i□ L, Pi□□□c□ S, Li□i□□ K, Paga□i l, Ma□□□□ a□i□ K, O□chi□i□□□a G, Che□ A, Pa□a□ia□□a□ K, La□d M, Ha□□e□L, Jeff□ie□ CD, De□□e□JC, B□a□bi□a E-M, Ka□□a□ KP, R□hde M, S□□i□g S, G□□e□M, W□□□e T, B□i□□□□ J, Ei□e□ JA, Ma□□□□i□□ V, H□ge□h□□□ P, K□□□ide□ NC, K□e□□ H-P, l□a□□□a N:** Complete genome sequence of *C□□□□□□□□□□□ □□□□□□* type strain (LIM-21). *□□□□□ G□□□□ □□ □□□* 2011, **4**:221-232.

88. **B□□□a□ JP:** Description of Cellulophaga algicola sp. nov., isolated from the surfaces of Antarctic algae, and reclassification of *C□□□□□□□□ □□□□□□□□* (ZoBell and Upham 1944) Reichenbach 1989 as *C□□□□□□□□□□ □□□□□□□□* comb. nov. *I□□J □□□□E□□□□ □□□□□□□* 2000, **50 P□5**:1861-1868.

89. **A□a□□ Rl, L□d□ig W, Sch□eife□ KH:** Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *□ □□□□□□□□ □ □□* 1995, **59**:143-169.

90. **Ha□de□□□a□ J, R□□d□□ MR, B□ad□ SF, C□a□d□ J, G□□d□a□ RM:** Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *C□□□ B□□□* 1998, **5**:R245-R249.

91. **Sch□□□ PD, Ha□de□□□a□ J:** Metagenomics for studying unculturable microorganisms: cutting the Gordian knot. *G□□□□ □ B□□□* 2005, **6**:229.

92. **P□□e PB, De□□a□ SE, J□□e□ M, T□i□ge SG, Ba□□□ K, Ma□fa□i SA, McHa□d□ AC, Che□g J-F, H□ge□h□□□ P, McS□ee□e□ CS, M□□i□□□ M:** Adaptation to herbivory by the Tammar wallaby includes bacterial and glycoside hydrolase

profiles different from other herbivores. □□□□ □ □□□ *A*□□□ □□□ □ □ *A* 2010, **107**:14793-14798.

93. **S□e□ G, Sc□□□ JJ, A□□□a□d FO, Ada□ □ SM, T□i□ge SG, Pi□□□-T□□ □□ AA, F□□□e□CE, Pa□□□ M, Wei□e□PJ, Ba□□□ KW, G□□d□i□ LA, B□□ffa□d P, Li L, O□□e□be□ge□ J, Ha□□i□□ TT, S□a□e□ SC, D□□□h□e TJ, C□□ie CR:** An insect herbivore microbiome with high plant biomass-degrading capacity. □□□□ *G*□□□□ 2010, **6**:e1001129.
94. **Tee□i□g H, Wa□d□a□□ J, L□□ ba□d□□T, Ba□e□ M, G□□c□□e□ FO:** TETRA: a web-service and a stand-alone program for the analysis and comparison of tetranucleotide usage patterns in DNA sequences. *B□C B*□□□□□□□□□ □□□□□ 2004, **5**:163.
95. **Webe□ M, Tee□i□g H, H□a□g S, Wa□d□a□□ J, Ka□□abg□ M, F□ch□ BM, K□i□d□□□h A, K□□c□□□ C, Wiche□□ A, Ge□d□□ G, A□a□□ R, G□□c□□e□ FO:** Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *I□□E J* 2011, **5**:918-928.
96. **Sa□dbe□g R, Wi□be□g G, B□□□de□ Cl, Ka□□e A, E□□be□g I, C□□□e□ J:** Capturing whole-genome characteristics in short sequences using a na□ve Bayesian classifier. *G*□□□□ □ □□□2001, **11**:1404-1409.
97. **McHa□d□ AC, Ma□□□ HG, T□i□ig□□ A, H□ge□h□□□ P, Rig□□□□□□ I:** Accurate phylogenetic classification of variable-length DNA fragments. □□□□ □□□□□□ 2007, **4**:63-72.
98. **B□ad□ A, Sa□be□g SL:** Phymm and PhymmBL: metagenomic phylogenetic classification with interpolated Markov models. □□□□ □□□□□□ 2009, **6**:673-676.
99. **Dia□ NN, K□a□□e L, G□e□□a□□ A, Nieha□□ K, Na□□e□□e□ TW:** TACOA: taxonomic classification of environmental genomic fragments using a kernelized nearest neighbor approach. *B□C B*□□□□□□□□□ □□□□□ 2009, **10**:56.
100. **Tee□i□g H, G□□c□□e□ FO:** Current opportunities and challenges in microbial metagenome analysis--a bioinformatic perspective. *B*□□□*B*□□□□□□□□ 2012, **13**:728-742.
101. **P□de□□S, Gaa□□e□a□d T, A□e□ EE:** A database of phylogenetically atypical genes in archaeal and bacterial genomes, identified using the DarkHorse algorithm. *B□C B*□□□□□□□□□ □□□□□ 2008, **9**:419.
102. **H□a□g S:** A bioinformatic pipeline for the taxonomic assignment of metagenome fragments for future analysis of an uncultured anaerobic planctomycete from a Black Sea microbial mat. *D*□□□□□ □□□□□□. University Bremen; 2009.
103. **Pe□e□□e□ TN, B□□□a□ S, □□□ Heij□e G, Nie□□e□ H:** SignalP 4.0: discriminating signal peptides from transmembrane regions. □□□□ □□□□□□ 2011, **8**:785-786.

104. **K□□gh A, La**rsson **B,** von **Heij□e G, S**onn **ha**mm **e**r **EL:** Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J* Mol *B*iol 2001, **305**:567-580.

105. **Pe□□□ha□e□ A, Pe□□□ha□e□ J, A**m **a**nn **R:** Fluorescence in situ hybridization and catalyzed reporter deposition for the identification of marine bacteria. *A*ppl *E*nviron Microbiol 2002, **68**:3094-3101.

106. **Ec□e□□ EM, Ba□□ ga□□□e□ M, H□be□ IM, Pe□□□ha□e□ J:** Grazing resistant freshwater bacteria profit from chitin and cell-wall derived organic carbon. *E*nvironmental Microbiology 2013, :n/a.

107. **F□□a□i K, M□□a□a N, M□□i□ Y, Ni□hiji□a T:** Distribution of heterotrophic nanoflagellates and their importance as the bacterial consumer in a eutrophic coastal seawater. *J*ournal of Oceanography 1996, **52**:399-407.

108. **Thi□g□□ad TF, Lig□e□□R:** Theoretical models for the control of bacterial growth rate, abundance, diversity and carbon demand. *A*quatic Microbial *E*cology 1997, **13**:19-27.

109. **McC□□□i□e MJ, We□he□bee R, Bacic A:** Subcellular location and composition of the wall and secreted extracellular sulphated polysaccharides/proteoglycans of the diatom Stauroneis amphioxys Gregory. Protoplasma 1999, **206**:188-200.

110. **Wa□e□□e□□ L, Bie□fai□A:** Localisation et r□le des [beta]-1,3-glucanes (callose et chrysolaminarine) dans le genre 'Pinnulari ' (diatom□es). La *C*ellule 1987, **74**:198-226.

111. **G□□de□□e□ JK, J□□ge□□e□ BB, La□□e□ E, Ja□□a□ch HW:** Mats of giant sulphur bacteria on deep-sea sediments due to fluctuating hydrothermal flow. Nature 1992, **360**:454-456.

112. **Scha□e□ R, R**øy **H, A**u**g**ust ¯n **N, Ge□□e□ich HH, Pe□e□□ M, We□□h□efe□ F, A**m **a**nn **R, Me□e□die□□□ A:** Bacterial sulfur cycling shapes microbial communities in surface sediments of an ultramafic hydrothermal vent field. *E*nviron Microbiol 2011, **13**:2633-2648.

113. **Ve□□□□a A, O□e□ A, a□d Ma Y:** *H*alophiles and hypersaline environments current research and future trends . Berlin: Springer; 2011.

114. **B**utt **MS, Tahi□-Nadee□ M, Ah□ad Z, S□□a□ MT:** Xylanases and their applications in baking industry. *F*ood Technology and *B*iotechnology 2008, **46**:22-31.

115. **G**uo **B, Che□ XL, S**un **CY, Zh□□ BC, Zha□g YZ:** Gene cloning, expression and characterization of a new cold-active and salt-tolerant endo-beta-1,4-xylanase from marine *G*laciecola mesophila KMM 241. *A*ppl Microbiol *B*iotechnol 2009, **84**:1107-1115.

116. **H□□g KS, Li□ SM, Fa□g TY, T□□□ WS, Li□ FP, S□□ KH, Ta□g SJ:** Characterization of a salt-tolerant xylanase from □□□□□ □□□□□□□□□□□□□□□□□ NTOU1. *B□□□□□□□□□□□□* 2011, **33**:1441-1447.

117. **A□de□□□□ I, Ti□da□□ BJ, P□□ □e□□e H, G□□e□ M, La□id□□ A, N□a□ M, C□□e□a□d A, G□a□i□a De□Ri□ T, Che□ F, Tice H, Che□g J-FF, L□ca□ S, Che□□□□□ O, B□□ce D, B□e□i□ T, De□□e□ JC, Ha□ C, G□□d□i□ L, La□d M, Ha□□e□L, Cha□g Y-JJ, Jeff□ie□CD, Pi□□□c□S, Pa□i A, Ma□□□□ a□i□ K, I□a□□□a N, O□chi□□i□□□a G, Che□ A, Pa□a□ia□□a□ K, Chai□ P, R□hde M, B□i□□□□ J, Ei□e□ JA, Ma□□□□i□□ V, H□ge□h□□□□ P, K□□□ide□ NC, K□e□□ H-PP:** Complete genome sequence of *H□□□□□□□□□□□* □□□□□□□□□□ type strain (AX-2). □□□□□□ *G□□□□□* □□ □□□2009, **1**:218-225.

118. **A□de□□□□ I, Sche□□e□C, G□□e□M, Ma□□□□ a□i□ K, H□□□e□SD, P□□a□l, K□e□□ H-PP, I□a□□□a N, K□□□ide□ N:** Novel insights into the diversity of catabolic metabolism from ten haloarchaeal genomes. □□□□ □ □□ 2011, **6**:e20237.

119. **Ta□g K, Jia□ N, Li□ K, Zha□g Y, Li S:** Distribution and functions of TonB-dependent transporters in marine bacteria and environments: implications for dissolved organic matter utilization. □□□□ □ □□ 2012, **7**:e41204.

120. **Ba□e□ a□ A, B□c□□f□M:** The structure of a LysM domain from E. coli membrane-bound lytic murein transglycosylase D (MltD). *J □ □□B□□□* 2000, **299**:1113-1119.

121. **G□□de□fe□d N, Kada□□ff LP:** Simple Lessons from Complexity. □□□□□□□ 1999, **284**:87-89.

122. **Jia□ N, Zhe□g Q:** The microbial carbon pump: from genes to ecosystems. *A□□□ E□□□□□□ □ □□□□□□□□* 2011, **77**:7439-7444.

123. **J□□c□□a M, P□□□ J, D□□ T, Q□e□□□E, P□a□a□ A:** Identification of active site carboxylic residues in *B□□□□□□* □□□□□□□□□□ 1,3-1,4-beta-D-glucan 4-glucanohydrolase by site-directed mutagenesis. *J B□□□C□□□* 1994, **269**:14530-14535.

124. **Ba□□ SG, M□□e□□ MK:** From bacterial glycogen to starch: understanding the biogenesis of the plant starch granule. *A□□□ □ □□ □□□□□B□□□* 2003, **54**:207-233.

125. **G□□□□e□JM, Fe□□□□de□-G□□ e□B, Fe□□□□de□-G□e□a A, G□□ e□-C□□□a□□a□ L, S□□che□ O, C□□-L□ad□ M, De□Ca□ □□ J, E□c□de□□ L, R□d□g□e□-Ma□□□□e□ R, A□□□□□-S□e□ L, La□a□a M, Pa□□e□ I, Neda□h□□□□□a□a O, Le□□□be□i I, Pi□ha□□i J, Ped□□□-A□i□ C:** Genome analysis of the proteorhodopsin-containing marine bacterium □□□□□□□□□□□□ sp. MED152 (*F□□□□□□□□□□□*). □□□□ □ □□□*A□□□ □□□□* □ *A* 2008, **105**:8724-8729.

126. **G□□i□□ JJ, W□e□e CR, S□a□e□ JT:** □□□□□□□□□□□ gen. nov., with three new species, □. □□□□□□□□ sp. nov., □. □□□□□□□ □□□□□ sp. nov. and □. □□□□□ □□□□□ sp. nov.,

gas vacuolate polar marine bacteria of the *C*-*F*-*B* group and reclassification of *F* as comb. nov. *I J B* 1998, **48 P 1**:223-235.

127. **Sac EL, a de Wie e PW, a de K ij D:** *F* as a model organism for characterizing biopolymer utilization in oligotrophic freshwater environments. *A E* 2011, **77**:6931-6938.
128. **Fe de -G e B, Rich e M, Sch e M, Pi ha i J, Aci a SG, G e JM, Ped -A i C:** Ecology of marine Bacteroidetes: a comparative genomics approach. *I E J* 2013, .
129. **Jia N, He d GJ, Ha e DA, Be e R, Ka e G, Wi he SW, Ki ch a DL, Wei ba e MG, L T, Che F, A a F:** Microbial production of recalcitrant dissolved organic matter: long-term carbon storage in the global ocean. 2010, **8**:593-599.
130. **O SP, a d Da T:** *A '* . Princeton: Princeton University Press; 2007.
131. **G c e FO, F ch BM, A a R:** Bacterioplankton compositions of lakes and oceans: a first comparison based on fluorescence in situ hybridization. *A E* 1999, **65**:3721-3726.
132. **Ki ch a DL:** The ecology of *C*-*F* in aquatic environments. *FE E* 2002, **39**:91-100.
133. **G e -Pe ei a PR, Sch e M, F ch BM, Be e C, Tee i g H, Wa d a J, Rich e M, Ba be V, Ba ai e E, G c e FO, A a R:** Genomic content of uncultured *B* from contrasting oceanic provinces in the North Atlantic Ocean. *E* 2012, **14**:52-66.
134. **G e -Pe ei a PR, F ch BM, A C, O i e MJ, a Be e JEE, A a R:** Distinct flavobacterial communities in contrasting water masses of the north Atlantic Ocean. *I E J* 2010, **4**:472-487.
135. **C e MT, Ki ch a DL:** Natural assemblages of marine proteobacteria and members of the *C*-*F* cluster consuming low- and high-molecular-weight dissolved organic matter. *A E* 2000, **66**:1692-1697.
136. **Che Y:** Bottom-up control on annually reoccurring bacterial community during spring bloom at station Kabeltonne (Helgoland). . Max Planck Institute for Marine Microbiology; 2012.
137. **I a a EP, A ee a YV, F a ie S, W igh JP, Zh a NV, G h a NM, Mi hai VV, Nic a DV, Ch i e R:** *F* gen. nov., sp. nov., a

novel member of the family *F*. *Int J Syst Evol Microbiol* 2004, **54**:705-711.

138. **Popper ZA, Michel G, Hervé C, Domozych DS, Willats WGT, Tuohy MG, Kloareg B, Stengel DB:** Evolution and diversity of plant cell walls: from algae to flowering plants. *Annu Rev Plant Biol* 2011, **62**:567-590.
139. **Aquino RS, Landeira-Fernandez AM, Valente AP, Andrade LR, Mourão PA:** Occurrence of sulfated galactans in marine angiosperms: evolutionary implications. *Glycobiology* 2005, **15**:11-20.
140. **Kreger DR, MEEUSE BJ:** X-ray diagrams of Euglena-paramylon, of the acid-insoluble glucan of yeast cell walls and of laminarin. *Biochim Biophys Acta* 1952, **9**:699.
141. **Chanzy AE, Sterte BA:** Structure of the paramylon from *E*. *Biochim Biophys Acta* 1960, **44**:161.
142. **Andersen RA:** Biology and systematics of heterokont and haptophyte algae. *Am J Bot* 2004, **91**:1508-1522.
143. **Chebli Y, Kaneda M, Zerzour R, Geitmann A:** The cell wall of the *Arabidopsis* pollen tube—spatial distribution, recycling, and network formation of polysaccharides. *Plant Physiol* 2012, **160**:1940-1955.
144. **Draget KI, Skjåk-Bræk O, Smidsrød O:** Alginates from Algae. In *Polysaccharides and Polyamides in the Food Industry*. Edited by Steinbüchel A, Rhee SK. Weinheim: Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim; 2005:1-30.
145. **Michel G, Tonon T, Scornet D, Cock JM, Kloareg B:** Central and storage carbon metabolism of the brown alga *E*: insights into the origin and evolution of storage carbohydrates in Eukaryotes. *New Phytol* 2010, **188**:67-81.
146. **Shin JH, Lee JH, Yi H, Cho J, Bae KS, Ahn TY, Kim SJ:** *K* gen. nov., sp. nov., an algicidal bacterium isolated from red tide. *Int J Syst Evol Microbiol* 2004, **54**:675-680.
147. **Cho JC, Giovannoni SJ:** Croceibacter atlanticus gen. nov., sp. nov., a novel marine bacterium in the family *F*. *Syst Appl Microbiol* 2003, **26**:76-83.
148. **Nedashkovskaya OI, Kim SB, Han SK, Rhee MS, Lysenko AM, Falsen E, Frolova GM, Mikhailov VV, Bae KS:** gen. nov., sp. nov., a novel member of the family *F* isolated from the green alga. *Int J Syst Evol Microbiol* 2004, **54**:119-123.

149. **Ch□i T-H, Lee HK, Lee K, Ch□ J-C:** Ulvibacter antarcticus sp. nov., isolated from Antarctic coastal seawater. *I*nternational *J*ournal of Systematic and *E*volutionary Microbiology 2007, **57**:2922-2925.

150. **Oh HM, Ka□g I, Fe□ie□a S, Gi**ov**a**nnon **i SJ, Ch□ JC:** Complete genome sequence of *C*roceibacter atlanticus HTCC2559T. *J B*acteriol 2010, **192**:4796-4797.

151. **Lee HS, Ka□g SG, K**won **KK, Lee JH, Ki□ SJ:** Genome sequence of the algicidal bacterium *K*ordia algicida OT-1. *J B*acteriol 2011, **193**:4031-4032.

152. **Y**oon **J-H, Ka□g S-J, Lee C-H, Oh T-K:** *D*okdonia donghaensis gen. nov., sp. nov., isolated from sea water. *I*nt *J* Syst *E*vol Microbiol 2005, **55**:2323-2328.

153. **Pi□ha□□i J, B**owm**a**n **JP, Neda□h□□□□a□a Ol, Le□□□be□i I, G**om**e**z-**C**ons**a**rn**a**u **L, Ped□□□-A**i□ **C:** Leeuwenhoekiella blandensis sp. nov., a genome-sequenced marine member of the family Flavobacteriaceae. *I*nt *J* Syst *E*vol Microbiol 2006, **56**:1489-1493.

154. **Neda□h□□□□□a□a Ol, Ki□ SB, Va□ca□□e□□M, S□a□□ae□□C, L**ys**e**nko **AM, R□hde M, F**rolov **a GM, Zh□□□□a NV, Mi□hai□□□ VV, Bae KS, Oh HW, S□i□g□ J:** *F*ormosa agariphila sp. nov., a budding bacterium of the family *F*lavobacteriaceae isolated from marine environments, and emended description of the genus *F*ormosa . *I*nt *J* Syst *E*vol Microbiol 2006, **56**:161-167.

155. **Da□i□ CL:** Production of laminarinase and alginase by marine bacteria after starvation. *FE□ □* Microbiology Letters 1992, **86**:349-355.

156. **P□□□e AM, Phi□i□□ MJ, Pe**nny **D:** Prokaryote and eukaryote evolvability. *B*iosystems 2003, **69**:163-185.

157. **L**évêqu **e E, bac□□a□h□ S□efa□ J, Ha□e B, Be□a□bi A:** Thermophilic archaeal amylolytic enzymes. *E*nzyme and Microbial Technology 2000, **26**:3-14.

158. **Hei□e N, Si□gh D,** van **de□ We□H, Sa□□i SO, J**ohnson **JM, Fea□□e□ CL, K□e□e□CM, P□e□ia□□ JO, Me□d□□□a-P□e□ia□□ L, We□□CM:** Molecular analysis of a UDP-GlcNAc:polypeptide alpha-N-acetylglucosaminyltransferase implicated in the initiation of mucin-type O-glycosylation in Trypanosoma cruzi *G*lycobiology 2009, **19**:918-933.

159. **M□□e□ IE, Pe**ttol**¨**no **FA, Ha□□ C, La□ □□g□a□i ER, Wi□a□□ WG, Bacic A:** Glycan profiling of plant cell wall polymers using microarrays. *J* Vis *E*xp 2012, :e4238.

160. **Le□□□ic I, B**ork **P:** Interactive Tree Of Life (iTOL): an online tool for phylogenetic tree display and annotation. *B*ioinformatics 2007, **23**:127-128.

161. **Barbeyron T, Michel G, Potin P, Henrissat B, Kloareg B:** iota-Carrageenases constitute a novel family of glycoside hydrolases, unrelated to that of kappa-carrageenases. *J Biol Chem* 2000, **275**:35499-35505.

162. **Hehemann J-H:** Structural and functional organisation of the agarolytic enzyme system of the marine *Flavobacterium Zobellia galactanivorans. PhD thesis.* L'Université Pierre et Marie Curie; 2009.

6. Ac□□□□ □edge□ e□□□

I thank Prof. Rudi Amann for giving me the chance of doing research in his department. It is an amazing research community that I love to work in and he is the key part of building it. His analytic mind and ecologic thinking always remind me of not just sitting on the data but also the need to connect the dots. Also, I thank him for being my first supervisor and advisor.

I thank Prof. Karl-Heinz Blotevogel for his well-timed support. Without him, this thesis would be impossible. It is such a pleasure to talk with him about science and life.

I thank Prof. Ulrich Fischer for his long-term support. His humorous way of reflecting on science inspires me to find a better way of understanding the problems I encountered in my study. I am also grateful that he participated in my thesis committee and acted as my third supervisor.

I can□t express enough my gratitude towards Dr. Hanno Teeling. It was him who brought me to this field. Even after four years of working with him, I am still amazed by his non-depleted new ideas and his encyclopedic knowledge of our human universe.

I thank PD Dr. Bernhard Fuchs for his contribution to my thesis. He is cautious about making statements and it reminds me that scientist has a big burden of proof. We make hypotheses, but we also need to try our best to prove them.

I also thank Prof. Frank Oliver Gl□ckner. I still remember it was him who took my call and explained bioinformatics to me in the middle of the day. I also remember it was him who redirected my r□sum□ to Hanno so that I finally found a place in this institute. It was him who provided the infrastructure that I needed to finish my study here.

I also thank the people that I worked closely with: Jost Waldmann, Marc Weber, Alexander Mann, Johannes Werner, Tobin Hammer, Gabrielle Moraru and Hikaru Suenaga. They provided me with many things that I could not accomplish myself alone.

There are also a lot of people who supported and help me along this long way. People from Molecolular Ecology Department like Kyoko Kubo, Christin Bennke, Anke Meyerdierks, Lizbeth Sayavedra, Fabian Ruhnau, Chia-I Huang, Regina Schauer, Lars Schreiber and Shi Yan, and people from the MGG such as Christian Quast, Elmar Pr□sse and Pelin Yilmaz. It is them who gave the cold science warm human faces. I thank my former and current office mates Pier Luigi Buttigieg, Basak Ozturk, Lennart Kappelmann and Peng Xing for their inspirational conversations.

6. Acknowledgements

I sincerely thank the French colleagues in Roscoff. Although it was only a five-weeks stay, they really made me feel at home. I feel forever in debt of Directors Mirjam Czjzek and Gurvan Michel. The "simple and happy" Tristan Barbeyron, the merry Lionel Cladière, the impressive Cécile Hervé and the high priest Murielle Jam, Elizabeth Ficko Blean, Robert Larocque, Jonas Collén are all in my "thank you" list. Their family-like atmosphere could break the ice and integrate any visitor in a second. I am also deeply impressed by their research and their attitude towards the truth.

There are also people who here and there, now and then helped me and supported me without their names here. I want to say I remember all your good deeds and merits. The only way to thank you is my helping people, like you have helped me.

I also thank my parents, my friends, teachers and colleagues in my motherland China. They taught me the Chinese ways of handling things and the Confucian ways of interacting with people. It was around them where I lay the foundation for my natural and social sciences.

It was an "eight-years battle" that now comes to an end. I feel really privileged in having this amazing journey with so many amazing people in the first half of my life. This experience will no doubt come with me all along. However, "it's not who I am that underneath, but what I do that defines me". And now it's time for me to unload my skill and contribute to the greater good.

7. A□□e□di□

7.1 The gill chamber epibiosis of deep-sea □□□ □□□□□ □□□□□□□□ shrimp thoroughly investigated by metagenomics and discovery of zetaproteobacterial epibionts

□

Authors

Cyrielle Jan[1], Jillian M. Petersen[2□], Johannes Werner[2,3□], Hanno Teeling[2□], Sixing Huang[2], Frank Oliver Gl□ckner[2,3], Olga V. Golyshina[4], Nicole Dubilier[2], Peter N. Golyshin[4], Mohamed Jebbar[1,5□] and Marie-Anne Cambon-Bonavita[6□*]

[1] Universit□ de Bretagne Occidentale, UMR 6197-Laboratoire de Microbiologie des Environnements Extr□mes (LM2E), Institut Universitaire Europ□en de la Mer (IUEM), Place Nicolas Copernic, 29 280 Plouzan□, France
[2] Max Planck Institute for Marine Microbiology, Celsisusstra□e 1, 28359 Bremen, Germany
[3] Jacobs University Bremen gGmbH, Campus Ring 1, 28 759 Bremen, Germany
[4] School of Biological Sciences, Bangor University, Bangor, Gwynedd, Wales, UK
[5] CNRS, LMEE, Technop□le Pointe du diable, BP70, 29 280 Plouzan□, France
[6] Ifremer, Centre de Brest, REM/EEP/LM2E, Technop□le Brest-Iroise, BP70, 29 280 Plouzan□, France

□ These authors contributed equally to this work.

*Corresponding author: Marie-Anne Cambon-Bonavita

Publication status

In preparation

My contribution

I assisted in taxonomic classification and functional annotation of the metagenomic data. Once the taxobins were formed, I generated the CAZyme profiles for the major bacterial clades and investigated their carbohydrate metabolic potential. I tried to connect the metabolic potential with ecological roles. I also participated in the interpretation and discussion of the results. I took part in the revision of the manuscript.

□ □

Abstract

The shrimp [illegible] thrives on deep-sea hydrothermal chimneys along the Mid-Atlantic Ridge and harbors a dense community of epibiotic bacteria in its gill chamber, dominated by filamentous *E*-[illegible]- and *G*[illegible]. Using metagenomics on shrimp specimens from the Rainbow hydrothermal vent field, we showed that both of these epibionts are chemolithoautotrophs that are able to use reduced sulfur compounds and hydrogen as electron donors, and oxygen and nitrate as electron acceptors. However, these episymbionts are distinct as, for example, they employ different CO_2 fixation pathways (*E*[illegible]: reductive tricarboxylic acid cycle; *G*[illegible]: Calvin-Benson-Bassham cycle) and differ in their ability for endogenous ammonium production. Such differences likely enable these epibionts to avoid direct competition with one another and at the same time enhance their host's fitness by allowing it to cope with a broader range of environmental conditions in its dynamic habitat. Furthermore, we identified genes that indicate possible molecular mechanisms of shrimp-epibiont interactions, as well as genes that provide nutritional and detoxification processes to the shrimp that improve its fitness. Besides the two main symbionts, the metagenome also contained sequences affiliated with [illegible], which could explain the presence of iron oxyhydroxide deposits in [illegible] mouthparts. We confirmed presence of [illegible] by fluorescence [illegible] hybridization and could thereby provide the first evidence for a [illegible]-invertebrate association.

Keywords

bacterial epibionts / gill chamber / hydrothermal vent / metagenome / [illegible] / [illegible]

7.2 Complete genome sequence of the algae-associated marine flavobacterium *F□□□ □□□ □□□□□□□□□□* KMM 3901^T

Authors

Alexander J. Mann[1,2], Richard Hahnke[1], Sixing Huang[1], Johannes Werner[1,2], Tristan Barbeyron[3], Bruno H□ttel[4], Kurt St□ber[4], Richard Reinhardt[4], Bernhard M. Fuchs[1], Jens Harder[1], Frank Oliver Gl□ckner[1,2], Rudolf I. Amann[1□], Hanno Teeling[1]

[1] Max Planck Institute for Marine Microbiology, Celsiusstra□e 1, 28359 Bremen, Germany
[2] Jacobs University Bremen gGmbH, Campus Ring 1, 28759 Bremen, Germany
[3] UPMC University Paris 6, UMR 7139 Marine Plants and Biomolecules, Station Biologique de Roscoff, 29682 Roscoff, Bretagne, France
[□□] Max Planck Genome Centre Cologne, Carl-von-Linn□-Weg 10, 50829 Cologne, Germany□

□ Corresponding author: Rudolf I Amann

Publication status

In preparation

My contribution

I assisted in the generation and validation of the CAZyme profile for *F□□□ □□□ □□□□□□□□□□* KMM 3901^T. I further explored the genomic potential of *F. □□□□□□□□□□* KMM 3901^T in the degradation of alginate, agar, laminarin, hemicellulose and □-glucan. I took part in the characterization of the PULs and substrates. I also assisted in the characterization of the metabolic pathways in glycolysis, fermentation and nitrogen recycling. I participated in the comparison among *F. □□□□□□□□□□* KMM 3901^T, *F□□□ □□□* group A and group B. Finally, I discussed with the other authors about the ecological niche of *F. □□□□□□□□□□* KMM 3901^T and its dependence on algae.

Abstract

In recent years, representatives of the marine *Bacteroidetes* have been increasingly recognized as specialists for the degradation of macromolecules. *Formosa* constitutes a *Bacteroidetes* genus within the class *Flavobacteriia*, whose members have been found where high levels of organic matter predominate, such as in association with marine algae, invertebrate animals and fish feces. So far, no *Formosa* representative has been sequenced. Here we report on the generation, annotation and analysis of the genome of the *Formosa agariphila* type strain (KMM 3901^T). This genome revealed a high level of metabolic flexibility, and is characterized by 13 polysaccharide utilization loci. The latter are large operon-like structures that comprise genes for TonB-dependent receptors, SusD proteins, glycoside hydrolases and oftentimes sulfatases, and play a crucial role in the biodegradation of complex polysaccharides.

Keywords

Formosa, polysaccharide utilization loci, carbohydrate-active enzymes (Cazymes)

7.3 Geomicrobiology of Hot Lake, a shallow-sea hydrothermal vent site off Panarea Island, Italy

Authors

Chia-I Huang[1], Rudolf Amann[1], Jan Amend[4], Wolfgang Bach[5], Jasmine Berg[3], Kai Uwe Hinrichs[2], Sixing Huang[1], Francesco Italiano[8], J□rg Peplies[1,6], Roy Price[3], Alban Ramette[1,7], Florence Schubotz[2], Roger Summons[3], Anke Meyerdierks[1*]

[1] Max Planck Institute for Marine Microbiology, Celsiusstrasse 1, D28359 Bremen, Germany
[2] MARUM, Center for Marine Environmental Sciences, Bremen, Germany
[3] Department of Earth, Atmospheric and Planetary Sciences, Massachusetts Institute of Technology, Cambridge, USA
[4] Washington University, St. Louis, USA
[5] University of Bremen, Department of Geosciences, Klagenfurter Strasse, D-28334 Bremen, Germany
[6] Ribocon GmbH, Fahrenheitstr. 1, 28359 Bremen, Germany
[7] Alfred-Wegener-Institut, Am Handelshafen 12, 27570 Bremerhaven, Germany
[8] Istituto Nazionale di Geofisica e Vulcanologia, Sezione di Palermo, Itaty

*Corresponding author: Anke Meyerdierks

Publication status

In preparation

My contribution

I assisted in the taxonomic classification and functional annotation of the metagenomes. I assessed the biodiversity based on rarefaction curves and Shannon indices and concluded that both samples were similarly diverse although their sample sizes were different. I identified sulfatases and other Pfam domains to get an overview about the microbial activities in the samples. Afterwards, I generated the CAZyme profiles for the two metagenomes and focused on the thermal stable □-glucanase family GH57. I investigated the GH57 family in several *E*□□□□□□□□□□□□□□□□□□□□ and concluded that this family is common in this CAZyme-poor proteobacteria class. Based on the abundances of *E*□□□□□□□□□□□□□□□□□□□□ in the metagenomes, I proposed that although □-glucans were common substrates in the shallow-sea hydrothermal vents, they were recycled through different sets of CAZymes. I also assessed the organic metabolic potential in the metagenome.

Abstract

Shallow sea hydrothermal vents are the habitat of unique microbial communities. In contrast to deep-sea vents energy can be gained by chemosynthesis as well as photosynthesis. An addition to carbon fixation, most deep sea vent systems are based on, allochthonous organic matter can contribute significantly to carbon assimilation at shallow vents. Although easier to access the biogeochemistry and microbial ecology of the more complex shallow sea systems has poorly been studied. Here, we investigated how reduced vent fluids impact the composition and function of microbial communities at Hot Lake, a strongly hydrothermally influenced, sediment filled depression 18 m below sea level off Panarea Island (Sicily, Italy). The temperatures measured at 10 cm sediment depth reached from 34[illegible]C to 74[illegible]C, correlating with distinctly different geochemical profiles, as well as with decreasing microbial cell counts and a changing microbial community composition with sediment depth and [illegible] temperature. Thermodynamic modelling based on pore water and fluid data revealed sulfur oxidation and sulfur reduction as the most favourable energy gaining processes at Hot Lake. This was supported by comparative 16S rRNA gene analysis and metagenome analysis which indicated [illegible] % of sequences affiliated with *E*[illegible], [illegible] % [illegible] *D*[illegible] and [illegible] % sequences related to the anaerobic phototrophic genus *C*[illegible] in the surface sediments. Cultured relatives of the detected species are mostly able to catalyze sulfur related metabolism, including sulfide oxidation, sulfur reduction or sulfate reduction. Metagenome analysis supported the relevance of sulfur metabolism and of the rTCA cycle for autotropic life at Hot Lake.

Keywords

shallow-sea hydrothermal vent, sediment, geochemistry, metagenomics, microbial community

Name: Sixing Huang Ort, Datum: Bremen, 11.03.2013

Anschrift: Niedersachsendamm 58, 28201 Bremen

ERKL□RUNG

Hiermit erkl□re ich, dass ich die Arbeit mit dem Titel:

A□□e□□i□g □he P□□□□accha□ide Deg□ada□i□□ P□□e□□ia□i□ Ma□i□e Mic□□bia□Ge□□□ e□a□d Me□age□□□ e□

selbstst□ndig verfasst und geschrieben habe und au□er den angegebenen Quellen keine weiteren Hilfsmittel verwendet habe.

Ebenfalls erkl□re ich hiermit eidesstaatlich, dass es sich bei den von mir abgegebenen Arbeiten um drei identische Exemplare handelt.

.....

(Unterschrift)