RNA based research - development, application and analysis within the MIMAS project

by

Dipl. Biol. Anna Klindworth

A thesis submitted in partial fulfilment of requirements for the degree of

DOCTOR OF PHILOSOPHY

in Biology

Approved Thesis Committee:

Prof. Dr. Frank Oliver Glöckner (chair)

Max Planck Institute for Marine Microbiology, Bremen, Germany

Jacobs University, Bremen, Germany

Prof. Dr. Matthias Ullrich

Jacobs University, Bremen, Germany

Prof. Dr. Jack Gilbert

University of Chicago, USA

Argonne National Laboratory

Date of Defense: October, 08th 2012

School of Engineering and Science

Statement of Sources

DECLARATION

I declare that this thesis is my own work and has not been submitted in any form for another degree or diploma at any university or other institution of tertiary education. Information derived from published or unpublished scientifc work has been cited in the text and listed in the references.

Signature **Date**

Thesis abstract

Every prokaryotic cell contains different sorts of ribonucleic acid (RNA) molecules, which are mainly dedicated to processing and regulating gene expression. The wide palette of functions is reflected by distinct RNA types, and each of them is represented by a complex and comprehensive research field. In particular, different culture-independent applications have attracted attention recently because the majority of microbes still resists cultivation until today. For example, the standard approach for microbial diversity studies is based on the comparative analysis of the evolutionarily conserved 16S ribosomal RNA gene (16S rDNA), and messenger RNA (mRNA) based metatranscriptomics allows culture independent gene expression analysis without prior knowledge of the present microbes or transcripts.

16S rDNA biodiversity studies, metatranscriptomics and other 'omic' applications play a central role within the MIMAS project, which aims at characterizing a bacterioplankton community at the long-term ecological research site Helgoland Roads. However, culture-independent applications have their limitations, and a careful design of experimental procedures is crucial to assure that these limitations do not overtly bias the results. Therefore, this thesis outlines the development and application of an improved pipeline for the analysis of metatranscriptomic data and the evaluation of PCR primers used to amplify 16S rRNA. In particular, the outcome of the latter serves as a guideline for enhanced research to find the most suitable primer pair for 16S rDNA biodiversity analysis in any habitat using any currently available sequencing technology.

The methods developed were used in a multi 'omic' study to characterize the phylogenetic and functional potential of the microbial community. The results identified the key players of an observed bacterioplankton bloom at Helgoland Roads and provided the first insights into taxonomically distinct nutrient strategies. They indicated that *Flavobacteria, Gammaproteobacteria* and *Alphaproteobacteria* are specialized for successive degradation of different algal primary products. This provided a series of ecological niches, allowing certain community members to grow. The results helped to uncover the secret of how members of the bacterioplankton can evade extinction despite the limited resources in the habitat. The work accomplished allows future follow-up studies and furnishes scientific society with guidelines to perform accurate diversity studies. Moreover the outcome serves as basis for future ecosystem monitoring.

Table of contents

List of abbreviations

ABC	ATP binding cassette
ATP	adenosin triphosphate
BAH	Biologische Anstalt Helgoland
bp	base pairs
CARD-FISH	catalyzed reporter deposition fluorescent-*in situ*-hybridization
CAZY	carbohydrate-active enzymes
CBM	carbohydrate-binding-module
CCD	charge-coupled device
cDNA	complementary deoxyribonucleic acid
CFG	*Cytophaga-Flavobacterium-Bacteroides*
CRISPR	clustered regularly interspaced short palindromic repeats
DNA	deoxyribonucleic acid
dNTP	deoxyribonucleoside triphosphates
EMAU	Ernst-Moritz-Arndt-University
emPCR	Emulsion polymerase chain reaction
FISH	fluorescence-*in-situ*-hybridization
GH	Glycoside Hydrolases
GOS	global ocean sampling
hr	hour
HV	hypervariable
IMaB	Institute of Marine Biotechnology
IOW	Leibniz Institute for Baltic Sea Research
KEGG	Kyoto Encyclopaedia of Genes and Genomes
LSU	large subunit

LTER	long term ecological research side
LWL	low-molecular weight
MIAME	Minimum Information About a Microarray Experiment
MIGS	Minimum Information about a Genome Sequence
MIMARKS	Minimum Information about a MARKer Sequence
MIMAS	Microbial Interactions in MArine Systems
MIMS	Minimum Information about a Metagenome Sequence
MPI	Max Planck Institute
mRNA	messenger ribonucleic acid
NCBI	National Center for Biotechnology Information
NGS	next generation sequencing
NPP	net primary production
nt	nucleotide
PacBio	Pacific Bioscience
PCR	polymerase chain reaction
PPi	inorganic pyrophosphate
RNA	ribonucleic acid
rRNA	ribosomal ribonucleic acid
RT PCR	reverse transcription polymerase chain reaction
SBP	solute-binding proteins
SMRT	single molecular real time
sRNA	small regulatory RNA
sRNA	small ribonucleic acid
SSU	small subunit
SusD	starch utilization system protein
TBDT	TonB denpendent transport
tmRNA	transfer messenger RNA

TP	*german*: Teilprojekt
TRAP	tripartite ATP independent
tRNA	transfer ribonucleic acid
TTT	tripartite tricarboxylate transporters
UTR	untranslated region
ZMW	zero-mode waveguide

List of figures

List of tables

1. Chapter

Introduction

1.1. The multiple facets of the RNA

In 1953, Watson and Crick successfully unravelled the molecular structure of deoxyribonucleic acid (DNA) (1). After this ground breaking discovery, researchers focused on the structure of ribonucleic acid (RNA) as the next puzzle to be solved on the road to understanding the molecular basis of life (2). Although DNA, as the gene carrier, remained in the spotlight for many years, it soon became clear that RNA participate more actively in many functions of the cell than originally thought.

RNA exhibits a wide variety of types, and each molecule is involved in different functions and activities. Some organisms, for example retroviruses, use RNA instead of DNA as a gene carrier. Other RNA molecules are bi-functional, such as the bacterial transfer-messenger RNA (tmRNA) (3). However, the most commonly known types in prokaryotes are messenger RNA (mRNA), transfer RNA (tRNA), ribosomal RNA (rRNA) and small, regulatory RNA (sRNA). mRNA functions as a translator and is primarily composed of coding sequences carrying the genetic information. The mRNA molecule is a single-stranded molecule forming no base pairing. Besides coding segments, it also contains noncoding, or untranslated sequences that may carry instructions for how the mRNA is handled by the cell. The untranslated region at the 5' end of the mRNA molecules found in *Bacteria* and *Archaea* is described as the Shine-Dalgarno sequence, which is essential in the binding of the mRNA to ribosomes. With the exception of a few molecules (4,5), prokaryotic mRNA do not exhibit a PolyA tail. Instead, they bear a triphosphate at the 5’end and a stem-loop structure at the 3’end. Moreover, the majority of prokaryotic mRNAs are polycistronic.

Although mRNA exhibits a central role within cellular protein biosynthesis, two other RNA types are essential for a fully functional process: tRNA and rRNA. In prokaryotes, rRNA consists of three types: 5S, 16S and 23S. With the exception of some microbes such as *Planctomycetes* (6,7), all three molecules are co-organized in one operon structure and synthesized as a single transcript that must be processed and cleaved at specific sites.

Finally, the rRNA strands and ribosomal proteins assemble to functional ribosomes. In all organisms, mature ribosomes are composed of two subunits: the large subunit (LSU) and the small subunit (SSU). In prokaryotes, the 50S LSU comprises two rRNA strands (23S and 5S) and 31 proteins. The 30S SSU is composed of the 16S rRNA and 21 proteins. Within the translation process, mRNA is bound between the subunits, and the ribosome catalyses the formation of the peptide bond. tRNA molecules serve as molecular adaptors carrying amino acids to the growing polypeptide. All cells contain individual tRNAs for each amino acid, which share a similar overall structure. The molecule is approximately 80 nucleotides (nt) long and exhibits a characteristic cloverleaf structure that results from complementary base pairing between different regions of the molecule. Mature tRNAs fold into a compact L-like structure, which is likely required for ribosome binding during translation.

Among the triumvirate of tRNA, rRNA and mRNA, bacterial genomes also exhibit many, perhaps several hundred, genes encoding for small, regulatory RNAs (8). Unfortunately, the nomenclature for describing those molecules has been neither uniform nor entirely satisfactory (9). Recently, the abbreviations ‘small RNAs’ or sRNAs dominated the literature (10). Therefore, we will refer to sRNA in this study.

The sRNA molecules range from 50 to 250 nt in length (10) and can be generated via processing or as primary transcripts (11). Although they were first observed in *Escherichia coli* four decades ago (12,13), the function remained mysterious for a long time. Today, it is known that sRNAs are crucial regulators of the prokaryotic gene expression, mRNA degradation, adaptation and virulence (14). For example, they can repress translation through direct interaction with the mRNA or by blocking the ribosome binding site. Other constitutively expressed sRNAs, the so-called clustered regularly interspaced short palindromic repeats (CRISPR), play a crucial role in the antiviral defence system (15) which is possibly performed by an RNA-interference-like mechanism (16).

It is interesting to note, that an RNA molecule exhibits the same cellular function among the three domains of life but differs a lot in terms of sequence. mRNA sequences vary based on the translated gene. The nucleotide sequence of sRNA is coupled to its functions. A similar situation occurs for tRNA molecules, which appear to evolve rapidly. Recent studies showed that the main factors driving tRNA evolution are most likely duplications, deletions and horizontal gene transfers (17), unlike rRNA, whose sequence is evolutionary conserved and ubiquitous among the three domains. Therefore it has been chosen as a marker gene for phylogenetic biodiversity studies. This demonstrates the wide variety of RNA based research, which goes far beyond classical molecular cell biology.

In summary, most of the prokaryotic RNA molecules are dedicated to processing and the regulation of gene expression. Each type of RNA plays a crucial and regulatory role within the complex cellular RNA network and comes with a complex and comprehensive research field focusing on different aspects. This thesis will focus on the impact of 16S rRNA and mRNA based research with respect to diversity and gene expression analysis in prokaryotes.

1.2. 16S rDNA based research

1.2.1. 16S rDNA as a marker gene

Microbes are ubiquitous (18) and their habitats range from terrestrial (19,20) to marine (21,22) and within humans (23,24) and plants (25). They participate in global cycles of energy transfer, use a wide range of substrates and possess many unique metabolic pathways (18,23). Therefore, understanding patterns and function of microbial diversity is of particular importance. Unfortunately, microbes are invisible to the naked eye and in general scientists have relied on pure culture experiments for identification and characterization. However, generating pure cultures, in particular from the marine habitat, turned out to be extremely challenging and limited investigations for a very long time. It is estimated that between 90-99% of the microbes resisted cultivation (26), most likely because of their inability to grow as mono-cultures or under standard laboratory conditions (27). Even for the extensively studied habitats, such as the human distal gut, only 20-40% of the known bacterial population has been cultivated so far (28). To overcome this 'cultivation-barrier', culture-independent surveys have been developed, such as fluorescence-*in-situ*-hybridization (FISH), metagenomics and comparative analysis using marker genes. Until today, the majority has been based on the comparative analysis of rRNA, in particular the small subunit (SSU) or 16S rRNA gene (16S rDNA), as introduced by Carl Woese in 1987 (29). Phylogenetic analysis based on the 16S rDNA sequences provided a first insight in the unseen microbial world without prior cultivation, and it became possible to identify uncultured microorganisms in almost any habitat.

Compared to other maker genes, the 16S rDNA has a number of clear advantages with regard to diversity analysis. The evolutionary conserved gene sequence is about 1500 bp in length, which is long enough to provide distinguishing and statistically valid measurements (30). The nine hyper variable (HV) regions allow accurate taxonomic and phylogenetic identification of prokaryotes, and the highly conserved regions serve as ideal primer target positions. It is a ubiquitous gene and no lateral gene transfer seems to occur (31). In the past, the most commonly used approach has relied on the amplifying, cloning and sequencing of the 16S rDNA gene using universal PCR primers (26,32). Those classical clone library based studies have dominated the field for many years and have been applied to a variety of habitats (33-35). The principle steps are (a) extraction of genomic DNA, (b) amplification of the 16S

rDNA using universal primers, (c) generating clone libraries containing 16S rDNA fragments, (d) sequence determination from clones and (e) comparative analysis of the retrieved sequences (Figure 1). Sequence information is then submitted and stored within comprehensive databases such as SILVA (36), greengenes (37) or RDP II (38). Each of these databases accumulates 16S rDNA sequences, and provides them to the research community in different formats. The latest RDP-II Release 10.29 dating from 1st June 2012 contains 2,319,039 bacterial and archaeal 16S rDNA sequences in an aligned and annotated format. Greengenes offers SSU gene sequence alignment for browsing, blasting, probing, and downloading. The last update of greengenes was released in October 2011 containing 1,049,116 aligned 16S rDNA reads. SILVA provides the scientific community with comprehensive, quality-checked and regularly updated databases of aligned 16S and 18S rDNA sequences for all three domains of life. Recently, SILVA 111 has been released which contains 3,194,778 SSU and 288,717 LSU aligned sequences. The accumulated data in any of those databases is freely available and serves as a basis for comparative phylogenetic analysis. Furthermore, the sequence information can be transferred for the design of nucleic acid probes to be used in fluorescence-*in-situ*-hybridization (FISH) analysis (26). FISH is an integral part of the so-called ribosomal RNA approach (31), which allows quantitative diversity and location analysis. The combined application of cloning, sequencing and comparative analysis followed by FISH is commonly described as the full-cycle RNA approach (Figure 1), and provides a comprehensive approach for the phylogenetic characterization of diverse communities in a wide range of habitats.

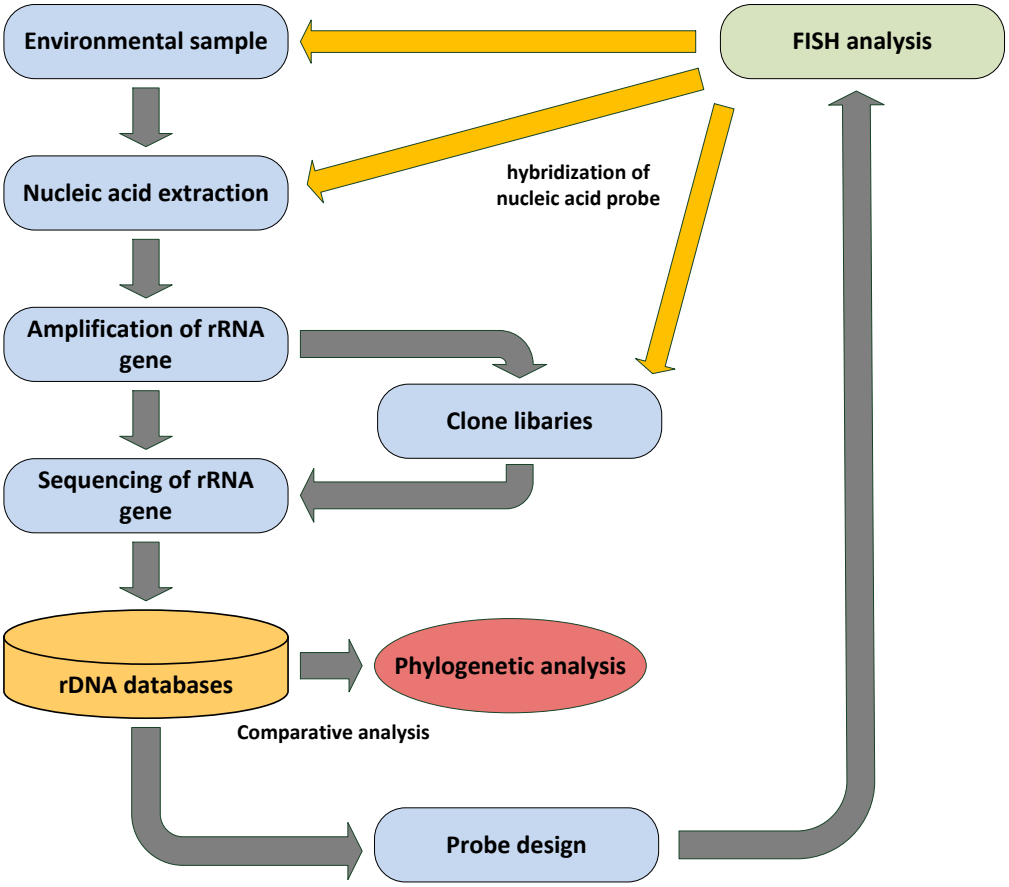


Figure 1: Schematic overview of the general workflow of the full cycle RNA approach.

1.2.2. Sequencing of the 16S rDNA gene

Sequencing of the 16S rDNA plays a central role within phylogenetic characterisation. In the past, 16S rDNA sequences from clone libraries were retrieved using the classical Sanger sequencing approach (39), which is still considered the 'gold standard' in terms of both read length and sequencing accuracy (40). However, it soon became clear that the microbial world remained dramatically under-sampled, indicating the need for new methods with higher resolution power. Emerging next generation sequencing (NGS) technologies appeared to be the answer to this request. These new approaches allowed the direct sequencing of PCR amplicons without the need for classical clone libraries but with a high throughput. Different NGS applications easily outperformed the classical Sanger approach by a factor of 100-1,000 on a daily basis, and reduced the cost per base in parallel (41). The intensive use of high throughput sequencing has increased the amount of 16S rDNA sequences in the databases with an enormous speed (42). Today, the most frequently used NGS technologies in microbial

biodiversity analysis include Roche's 454 pyrosequencing (43), Illumina (formerly Solexa) (44), Applied Biosystem's SOLiD (45-47) and Ion Torrent from Life Technologies (48). Details about the NGS methods are described in chapter 1.4.

In 2006, Roche's 454 became the first high throughput sequencing technology to be successfully applied in biodiversity analysis (22). For this purpose, the HV region six of the 16S rDNA was PCR-amplified and sequenced using the first generation of the 454 pyrosequencing platform (22). With the release of the 454 FLX and Titanium systems the throughput and resolution of 16S rDNA sequencing further improved (49). Consequently, Roche's 454 pyrosequencing technology has been used for microbial biodiversity analysis in a great range of different habitat types, such as soil (20), human (23,24) and marine (21,50) to name just a few. The continuous development further improved the technology in terms of accuracy, throughput and read length up to 1000 bp nowadays (49); hence, Roche's 454 pyrosequencing remains attractive even today.

Recently, other NGS technologies also made an advance in large scale diversity analysis. In particular, Illumina might supplant Roche's 454 pyrosequencing by offering a reduced per base cost and a higher sequencing depth (51). Despite having short read lengths, it has been successfully applied in a large variety of habitats such as marine (52), soil (53,54) and diverse human microbiomes (55-57). Ion Torrent sequencing, which has recently been used for diversity analysis in piggery waste treatment (58) and clinical research (59), satisfies with its relatively low cost and rapid sequencing speed. In fact, Life Technologies advertised that by the end of 2012 the system will be able to sequence the entire human genome in just a few hours (http://www.iontorrent.com/).

Despite the many advantages of high throughput sequencing approaches, the relatively short read lengths still possess a problem for in-depth phylogenetic analysis (32,42,60). Different HV regions exhibit varying degrees of sequence diversity, and no single HV region is able to distinguish among all bacteria (61). Therefore, phylogenetic accuracy is length dependent which is best achieved by full length sequences currently provided by classical clone libraries. Innovative results are expected from the fairly new high throughput 'single-molecule real-time' (SMRT) sequencing technology from Pacific Bioscience (PacBio) (62), which clearly benefits from its long read length up to 10,000 bp and an average of 3,000 bp. So far it has not been used for any published biodiversity studies but advertised advantages sound promising. Nevertheless, the other intensively used NGS technologies are still convincing, offering a snap-shot of the biodiversity with an extraordinary sequencing depth and using a fast and cost-effective method.

1.2.3. Limitations of PCR based 16S rDNA analysis

Several studies have proven the efficiency and suitability of the 16S rDNA as a marker gene; however, each application has its limitations. To begin with, the marker gene itself has to deal with some restrictions. In most cases a bacterial genome exhibits only one or two 16S rDNA copies (63), but some microbial cells contain multiple or heterogeneous amounts of 16S rDNAs, which are often associated with nucleotide sequence variability (64,65). For example, *Bacillus subtilis* has 10 copies (66) and *Clostridium paradoxum* up to 15 copies with heterogeneous intervening sequences (64). Those multiple copies can lead to an overestimation in terms of abundance and bacterial diversity (63). Furthermore, in some cases taxonomic resolution power can be insufficient at the species level or with closely related species (67,68).

Accurate PCR based 16S rDNA analysis heavily depends on the choice of primers (63). Using suboptimal primers can lead to under-representation (69) or selection against a single species or even whole groups (24,42,70). For example, the general primer 384F fails to detect *Verrucomicrobia* (71) and 967F matches only <5% of *Bacteriodetes* (22). Although a standard PCR is expected to tolerate up to one to two mismatches between the primer and its target (32), a primer mismatch might lead to preferential amplification, which results in a biased picture of the bacterial diversity (14, 41). Therefore, choosing the correct primer and especially primer pairs is a key element within PCR based diversity studies.

1.3. mRNA based gene expression analysis in prokaryotes

1.3.1. Metatranscriptomics in prokaryotes

Gene expression can be described as the cellular answer to an internal or external stimulus. The information encoded by each gene is used to synthesize a corresponding gene product. It allows the cell to adapt to environmental disturbances and its own changing need in a remarkably flexible and rapid way. The whole process itself is tightly regulated and controls the timing, volume and level of each individual gene. Analysis of cellular gene expression gains insights into the functional potential of microbes as well as the regulations, stimulation or inhibition of transcription.

In the past, several techniques for gene expression analysis have been developed and successfully established; for example, Northern Blotting and reverse transcription PCR (RT PCR). However, most approaches only allow investigations of single genes or just a few genes at a time. Development of cDNA based microarrays was among the first to settle the deficit; however, this technique is limited by the need of prior knowledge of the genes of interest. Thus, gene expression analysis of unknown diverse microbial communities remained challenging. With the advent of NGS technologies (see chapter 1.4) culture-independent studies of environmental samples became feasible. The so-called metatranscriptomic approach allows direct sequencing of cDNA without any cloning step or prior knowledge of the present genes (72). Briefly, the experimental set up of a prokaryotic metatranscriptomic pipeline includes a) extraction of total RNA, b) capture of mRNA, c) cDNA synthesis, d) sequencing of transcripts and e) data analysis (see also Figure 2). Coupled with NGS a great sequencing depth can be achieved, revealing the less dominant but interesting transcripts among the most abundant ones. In particular, Roche's 454 pyrosequencing, which persuades with longer a read length, has been widely used in the early stage of metatranscriptomics. The first study using this platform was published in 2006 (73). For the first time, it was possible to analyse the metatranscriptome of a mixed bacterial and archaeal soil community with a high sequencing depth. In the following years, other studies confirmed how Roche's 454 pyrosequencing can be applied with ease to access the unknown mRNA pool of microbial communities (74-76). For example, comparative day/night metatranscriptomic studies of microbial communities in the North Pacific subtropical gyre demonstrated the metabolic and

biogeochemical response of a bacterial community to solar forcing. During the day, transcripts of genes involved in energy processes such as photosynthesis, C1 metabolism and oxidative phosphorylation were more abundant. In contrast, genes encoding for proteins involved in housekeeping activities such as amino acid biosynthesis, vitamin biosynthesis and membrane synthesis and repair were highly expressed at night (77). This study successfully demonstrated how high-throughput sequencing technologies can be applied to analyse complex environmental metatranscriptomes of microbial communities without prior knowledge of what genes might be expressed.

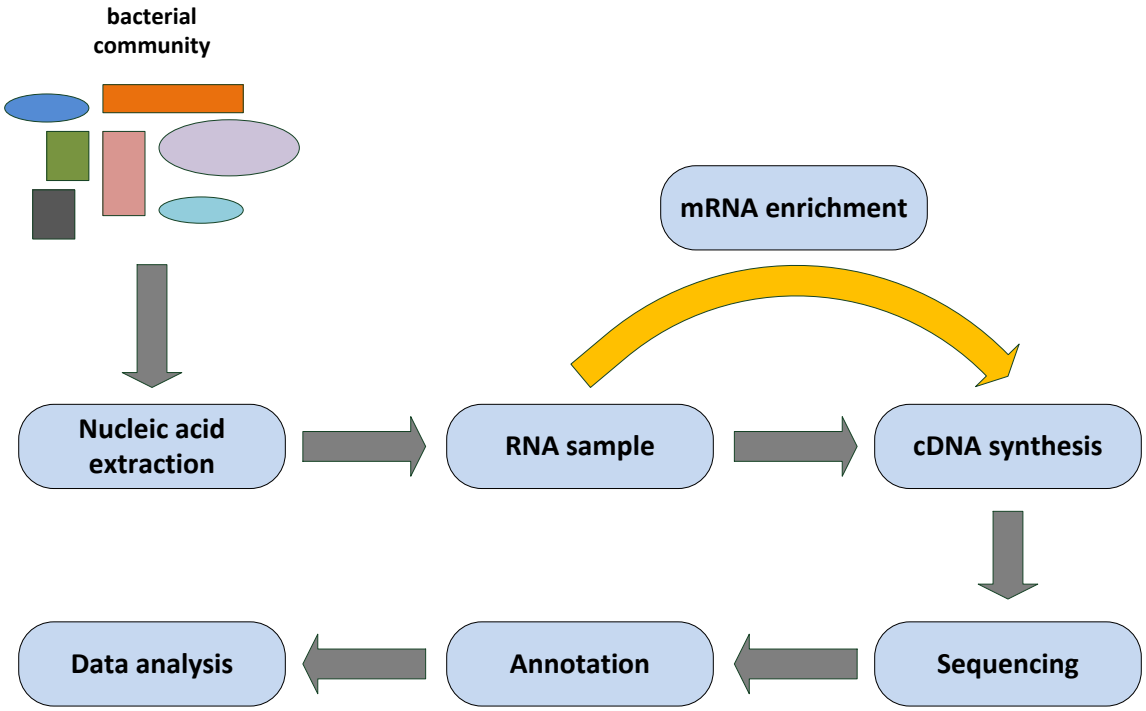


Figure 2: Metatranscriptomic analysis: Schematic overview of an experimental set-up.

At present, also Illumina (78,79) and SOLiD (80) are attracting attention and have been applied in a few prokaryotic studies, and hence, have demonstrated their potential despite short read lengths. No matter what NGS platform has been used, metatranscriptomics coupled with high throughput sequencing has clearly revolutionized environmental genomics. This technique not only allows culture-independent investigations but also analysis of the functional potential combined with accurate taxonomic resolution. Consequently, identification of the functional potential as well as the active members within different types of habitats such as marine (75-77,81-84), human (85) and soil (74,86) became feasible.

Another key advantage of NGS based gene expression analysis is the potential discovery of new genes, reannotation and the detection of untranslated regions (UTRs)(76,87,88). In particular detection of sRNAs has attracted a great deal of attention recently (8,10,85,89,90).
Metatranscriptomics enables the scientific community to identify regulation, adaptation and the level of gene expression of microbes in their natural environment. It is one of the most powerful tools for understanding the regulation and timing of complex microbial gene expression patterns in response to changing conditions and can also be seen as a potential tool for the discovery of novel biocatalysts with biotechnological applications (91).

1.3.2. Limitations of metatranscriptomics

More than 40 years ago, when mRNA was discovered, its defining characteristic was instability (92,93), which is nowadays an important parameter for the level and adaptation of gene expression in response to environmental disturbances (94). However, the short half-life of prokaryotic mRNA can also hamper the analysis (95-97), because their decay is frequently initiated shortly after or, in peculiar cases, even before their transcription is completed (98). In general, working with RNA always has to be well planned to avoid cellular stress and long handling times. Otherwise, a bias in the gene expression pattern might occur. Moreover, contamination with RNases that are found in a variety of environmental sources can be a potential risk for mRNA degradation. Thus, a RNase-free working equipment and atmosphere are mandatory.
Other challenges arise due to the composition of a total RNA sample, which contains approximately 95% rRNA and tRNA (99). This indicates the need for sensitive mRNA detection methods or prior removal of non-protein-coding RNAs. In eukaryotes, the capture of mRNA transcripts coupled with cDNA synthesis can be easily performed using Oligo(dT) primer (100). In prokaryotes, the common mRNA molecule lacks a 3′-end poly(A) tail. Thus, alternative methods are necessary to remove rRNA and tRNA before sequencing. There are several techniques available such as Oligo(dT) priming from artificial polyadenylated mRNAs (75,89), subtractive hybridization with rRNA-specific probes (101,102), reverse transcription with rRNA-specific primers followed by RNase H digestion to degrade rRNA:DNA hybrids (103), size separation and isolation of mRNA via gel electrophorese (104) and exonuclease digestion of rRNA molecules (105). Some methods have also been applied in a combined experimental pipeline, for example, subtractive hybridization and exonuclease digestion (106).

Successful mRNA capture is usually followed by cDNA synthesis using random hexamers. Although, this application is utilized to generate reads across the entire length of all expressed transcripts, it results in a bias in the nucleotide composition at the start of sequencing reads (107). On the contrary using Oligo(dT) priming (75), which is the common approach in eukaryotes (108,109), appears to be highly biased towards the 3'-end of the transcripts (107). In summary, each of these methods has its strengths and weaknesses and can always introduce a potential bias. However, no standardized experimental procedure has been accepted. Therefore, choosing the most appropriate method is a crucial step within the experimental set up.

Challenges also remain with respect to comparative metatranscriptomics. As previously described, there is no standardized protocol with respect to the available mRNA processing and sequencing technology. The experimental procedure can influence the mRNA output, and the applied NGS platform has a high impact on the quality and quantity on the data output. Another bias might result from the data processing. In this context, standards regarding data processing, the experimental procedure and the sequencing technology are strongly required to ensure comparable data among different studies (110,111). Such guidelines could include, for example, contextual data of the sampling point or habitat, and instructions on the data processing, including annotation, statistical evaluation, taxonomic classification and availability of sequences in online databases. Similar guidelines have been proposed for other research fields such as MIAME for microarrays (112), MIGS/MIMS for genome- and metagenome sequences (113) and MIMARKS for marker gene sequences (114). At present, metatranscriptomics is still at an early stage of research; thus generating standards is of particular importance and can influence the future research immensely.

It also has to be noted that metatranscriptomics do not reflect all regulatory processes in the bacterial cell such as post-transcriptional, translational and post-translational regulation (115,116). Although this is not a bias introduced by an experimental or technical issue, one has to keep in mind that the method itself has its limitations. If further information about mature gene products are requested, (meta)proteomics would be one method of choice (117). However, putting aside all limitations and challenges, metatranscriptomics is a promising research tool and is expected to gain even higher resolution power with future technical developments.

1.4. DNA Sequencing technologies

1.4.1. Next generation sequencing technologies

Sequencing technologies in general play a central role in a broad range of applications such as molecular cloning, whole genome analysis, transcriptomics and biodiversity studies. The permanent development and rigorous research of DNA sequencing tools in the past 30 years has truly changed our understanding of the molecular world.

The chain-termination method published in 1977 (39), also commonly referred to as the Sanger method, has remained the most commonly used DNA sequencing technique to date. It is still considered the 'gold standard' with regard to sequencing length and quality, and it has been applied in a multitude of projects, including the sequencing of the human genome (118) and the global ocean sampling (GOS) dataset (119). However, several new so-called next generation sequencing (NGS) methods have been developed in the last few years (Table 1). In particular, Roche's 454 pyrosequencing (43) and Illumina (44) found general approval from the scientific community. However, SOLiD (45-47) and Ion Torrent (48) also appealed to customers. In addition, the fairly new single molecular real time (SMRT) sequencing technology from PacBio (62) attracted a great deal of attention with regard to sequencing length. Although they all differ in terms of sequencing chemistry they share one common feature: 'high throughput sequencing'. This means that a single experiment or a sequencing run produces far more reads that the 96 or 384 well-based Sanger technology. In particular, increasing throughput associated with low cost per base has revolutionized DNA sequencing making it possible for even single research groups to generate large amounts of sequence data very rapidly and at a substantially lower cost.

In 2005, Roche's 454 pyrosequencing method was the first new sequencing platform available on the market. In contrast to the Sanger technology, it is based on a non-electrophoretic bioluminescence method. It measures the release of inorganic pyrophosphate by proportionally converting it into light using a series of enzymatic reactions (120). To begin with, emulsion PCR (emPCR) is used to prepare sequencing templates in a cell-free system. For this purpose, the DNA is sheared and oligonucleotide adaptors containing universal primer sides are ligated to the target ends, allowing complex genomes to be amplified with common PCR primers (72,120). After ligation, the DNA strands are separated into single strands and attached to beads. This is performed under conditions that favour one DNA

fragment per bead. The DNA bead complex is placed in a reaction mixture containing an oil–aqueous emulsion, followed by emPCR. The emPCR reaction generates multiple copies of the same DNA sequence on the surface of each bead. Finally, each bead contains up to 1,000,000 copies of the originally attached DNA molecule. This is necessary to produce a detectable signal for the sequencing reaction (46). These amplified single molecules are captured in a picotiter plate (PTP) that holds a single bead in each of several million single wells. The pyrosequencing reaction mixture is added and the pyrophosphate-based sequencing is performed in parallel in each single well. Nucleotide incorporation results in the release of inorganic pyrophosphate (PPi). ATP sulfatase converts Adenosine 5´ phosphosulfate and PPi into ATP. The latter is used as a substrate of the enzyme Luciferase to generate light, which can be detected with a charge-coupled device (CCD) camera. The cycle is iteratively repeated for each of the four bases. The light emitted is directly proportional to the amount of incorporated nucleotides, and is only limited by the detector saturation. In general, this technology achieves 99.9% accuracy (45) but sequencing errors may occur due to the presence of homopolymers as repeats greater than five nucleotides cannot be quantitatively measured (121). Nevertheless, Roche's 454 pyrosequencing satisfies with long reads up to 1,000 bp associated with a fast and cost effective performance generating up to one million reads per run.

Shortly after the release of the 454 pyrosequencing platform, Illumina (formerly Solexa) offered a new emerging sequencing by synthesis (SBS) NGS technology using four fluorescently labelled nucleotides to sequence millions of DNA clusters on one flow cell surface in parallel. During the first step, the library containing fix adaptors is denatured to single strands, attached to the flow cell and cloned via bridge amplification to form clusters of clonal DNA fragments. After library splicing into single strands, the first sequencing cycle begins by adding four labelled reversible deoxynucleoside triphosphate (dNTP), primers, and DNA polymerase. The nucleotide label serves as a terminator for polymerization, so after each dNTP incorporation, the fluorescent dye is imaged to identify the base and then enzymatically cleaved to allow incorporation of the next nucleotide. The cycles are repeated to determine the sequence of bases within the DNA fragment. The end result is a highly accurate sequencing process with low error rates that yields approximately three billion reads per run. A weakness of the platform is that it tends to produce biased sequence coverage that occurs in AT-rich repetitive sequences (122,123). However, with respect to the enormous high throughput rate, Illumina is clearly on the rise despite its relatively short read length of up to 150 bp of the HiSeq 2000 System.

The SOLiD platform is based on massive parallel sequencing by ligation. The initial library construction involves an emPCR reaction analogue to the 454 pyrosequencing technology. Next, amplification products are transferred onto a 3' modified bead located on a glass surface where the actual sequencing process occurs. DNA sequencing involves five different primers that differ in their position on the adaptor template by one nucleotide. In the first round, the primers bind to the adapter sequences which are sticking to the glass bead. A set of four fluorescently labelled di-base probes compete for ligation to the sequencing primer and a CCD camera detects the di-base signal after ligation. Specificity of the di-base probe is achieved by interrogating every 1^{st} and 2^{nd} base in each ligation reaction (for example, dibase probe: CANNNNNN or GGNNNNNN). Multiple cycles of ligation, detection and cleavage are performed with the number of cycles representing the read length. Finally, the extension product consisting of primer and di-base elements is melted away and a new primer is added so that the whole process can be repeated. The strength of the SOLiD system is the large amount of data output, however, the sequences are only 35-75 bp in length.

Ion Torrent uses a chip based semiconductor technology with a simple sequencing chemistry that is based on a well-characterized natural biochemical process: the incorporation of nucleotides during DNA synthesis. In particular, the hydrogen ion, which is released as a by-product, plays a central role within the process. Briefly, if a nucleotide, for example, Adenin, is incorporated during the process, the released hydrogen ion changes the pH of the sequencing solution, which will be detected by the ion sensor. The remaining nucleotides will be washed away. If the next nucleotide that floods the chip is not a match, no voltage change will be recorded and no base will be called. Two identical bases on the DNA strand results in voltage doubling, which can easily be distinguished by the sensor. However, Ion Torrent technology struggles with homopolymer-associated errors, whose accuracy is even worse that Roche's 454 pyrosequencing (124). Moreover, with 100 bp on average and comparatively lower throughput (0.25-0.4 million reads/run) it might be difficult for Ion Torrent to compete with the other sequencing technologies.

The newly emerging PacBio SMRT platform benefits from its potential to produce a read length greater than 3,000 bp on average. The technology employs the natural DNA replication process and relies on newly developed SMRT cells, each containing thousands of zero-mode waveguides (ZMW). Those SMRT cells enable single molecule real-time observation of individual fluorophores against a dense background of labelled nucleotides while maintaining a high signal-to-noise ratio (62). Each single DNA replication process is performed in a ZMW chamber with an active polymerase immobilized at the bottom. For sequencing, the DNA

strand of interest and four distinguishable fluorescently labelled deoxyribonucleoside triphosphates (dNTPs) are added. Unlike other NGS methods, PacBio uses alternatively labelled phosholinked nucleotides whose fluorescent dye is attached to the phosphate chain rather than to the base. As the DNA polymerase naturally incorporates nucleotides, the phosphate chain is cleaved and the dye molecule released. The latter quickly diffuses out of the detection ensuring low background noise as the process repeats. The light emitted by fluorescence is detected by a state-of the-art optical system developed by PacBio. In the course of incorporation, completely natural long DNA fragments can be sequenced. In summary, the PacBio platform distinguishes with respect to read length enabling very flexible applications. Moreover, no additional PCR step is needed, which reduces the bias during sample preparation. However, a low accuracy of approximately 85% (125) and a comparatively low throughput (75,000 reads/run) dampens the promising expectations.

Each of these approaches has its strengths and weaknesses, and qualifies for different research questions. Although the NGS based technique is still considered a very young field, several studies proved the power and impact of this technology. Improvements with respect to speed, read length, cost and accuracy in the near future are expected to influence molecular genomics even more.

Table 1: Characteristics, advantages and mechanisms of different sequencing platforms.

Sequencing platform	**454 GS FLX+**	**HiSeq 2000**	**Ion Torrent Proton Sequencer**	**SOLiD 5500xl with microbeads**	**SMRT**	**Sanger**
Sequencing company	Roche	Illumina	Life Technologies	Applied Biosystems	Pacific Bioscience	-
Sequencing mechanism	pyrosequencing	sequencing by synthesis	semiconductor sequencing	ligation and two-base coding	single-molecule real-time	dideoxy chain termination
Average read lenght	700 bp	100 bp	100 bp	35-75	3,000 bp	400-900 bp
Max. read lenght	up to 1,000 bp	150 bp	200 bp	85 bp	10,000 bp	1,000
Reads/run	1 Million	3 Billion	0.25-0.4 Million	1,200-1,400 Million	75,000	-
Output data/run	0.7 Gb	600 Gb	0.01-1 Gb	180 Gb	0.1 Mb	1.9-84 Kb
Time/run	23 hours	11 days	2 hours	1-7 days	1 day	20 min up to 3 hours
Accuracy*	99.9%	98%	99%	99.9%	~85%	99.9%
Advantages	long read length associated with fast performance	high throughput	fast sequencing runs	high accuracy	long read length	high quality, low error rates and long read length
Disadvantages	error rates in terms of hompolymeres	relatively short reads	relatively short read length, higher error rates	relatively short reads	low throughout, high error rate	high cost and very low throughput in comparison with the other NGS technologies

* numbers have been published either in Koren et al. (125), Liu et al. (45) or on the corresponding web page of the sequencing company

1.5. The MIMAS project

1.5.1. Aims of the MIMAS project (the big picture)

In the context of ongoing global changes (for example, global warming and eutrophication) the '**M**icrobial **I**nteractions in **MA**rine **S**ystems' (**MIMAS**) project aimed at investigating seasonal changes in microbial communities at well-defined long-term ecological research sites (LTER). The project focused on the establishment of new molecular biological techniques for the determination and characterisation of marine microbial assemblages. Briefly, the core of the project is based on different 'omic' approaches in order to unravel the complexity of the metabolism and lifestyle of diverse marine microorganisms, including those that remain uncultivated: Metagenomics addresses the genetic potential of the bacterial community as a whole, and metatranscriptomics and metaproteomics will shed light onto the active fraction of genes.

This integrated approach provided new insights into the ecological role of marine bacterial communities and their response to environmental changes such as climate change or fluctuations in the availability of nutrients. In addition, FISH and 16S rDNA sequencing were performed to explore the biodiversity at the LTER of choice. The comprehensive analysis of the genomic diversity and activity of marine microorganisms is a key for a better understanding of climate changes and other natural or anthropogenic influences (for example, eutrophication) on biogeochemical nutrient cycles. The valuable output allows follow-up analysis with respect to finding potential gene candidates for biotechnology or medical use and serves as a basis for future ecosystem monitoring.

1.5.2. Project partner and contributions

In order to address different research areas, the MIMAS project was divided into five different subprojects (Figure 3), which will be explained briefly.

Subproject one (G*erman: Teilprojekt 1* (TP1)) was supervised by Prof. Dr. Thomas Schweder's group at the Ernst-Moritz-Arndt-University (EMAU) and the Institute of Marine Biotechnology (IMaB) in Greifswald, Germany, and addressed the (meta)proteome of marine model organisms and the LTER Helgoland Roads (see chapter 1.6). In the beginning, the main focus was on the development and establishment of a reproducible experimental pipeline in order to analyse the (meta)proteome of marine model organisms such as the

planctomycete *Rhodopirellula baltica* SH1^T (6) and bacterial communities. The latter was potentially difficult due to the complexity of the sample. Further, new software programs for automatic *de novo* peptide sequencing needed to be designed.

The second subproject (TP2) was supervised by Prof. Dr. Frank Oliver Glöckner at the Jacobs University and the Max Planck Institute (MPI) for Marine Microbiology in Bremen, Germany, and focused on (meta)transcriptomic analysis. As a start, gene expression analysis of the marine model organisms *R. baltica* SH1^T (6) was performed using whole genome microarrays (126) and Roche's 454 pyrosequencing to set up the experimental pipeline. To complement and extend the studies, metatranscriptome analysis at the LTER Helgoland Roads was set as one of the key elements. This approach provided a comprehensive insight into the flexible adaptation of the gene expression of the bacterial community due to seasonal changes in the environment.

The core of the third subproject (TP3), which was under the direction of Prof. Dr. Rudolf Amann (MPI Bremen, Germany), analysed the metagenome and microbial diversity analysis at the LTER Helgoland Roads. This project included the set-up of the weekly sampling process and performance of 16S rDNA pyrosequencing, FISH analysis and metagenomics. The construction of metagenome libraries from the picoplankton fraction also served as the backbone for interpreting the data of the entire project. This multivariate dataset comprising physical and chemical parameters, zoo- and phytoplankton and microbial diversity data provided first insights in the interactions and adaptation of the different trophic levels within the microbial food web as well as the functionality of the biological processes present.

The basis of the fourth subproject (TP4) was located at the Leibniz Institute for Baltic Sea Research (IOW) in Warnemünde, Germany. The project was under the direction of Prof. Dr. Klaus Jürgens and focused on the LTER Gotland Deep in the Baltic Sea. Unlike the North Sea, the deeper Baltic Sea water layers are usually anoxic, resulting in redox clines within the water body. Samples from this significant LTER near the chemoclines were analysed by applying different 'omic' approaches, and were compared to the results to chemical profiles from the sampling sites.

The last but very important fifth subproject (TP5) can be described as a technology platform providing the bioinformatic backbone for the different 'omic' approaches. The focus was on the development of databases and software for high throughput mass spectrometry and NGS sequencing technologies.

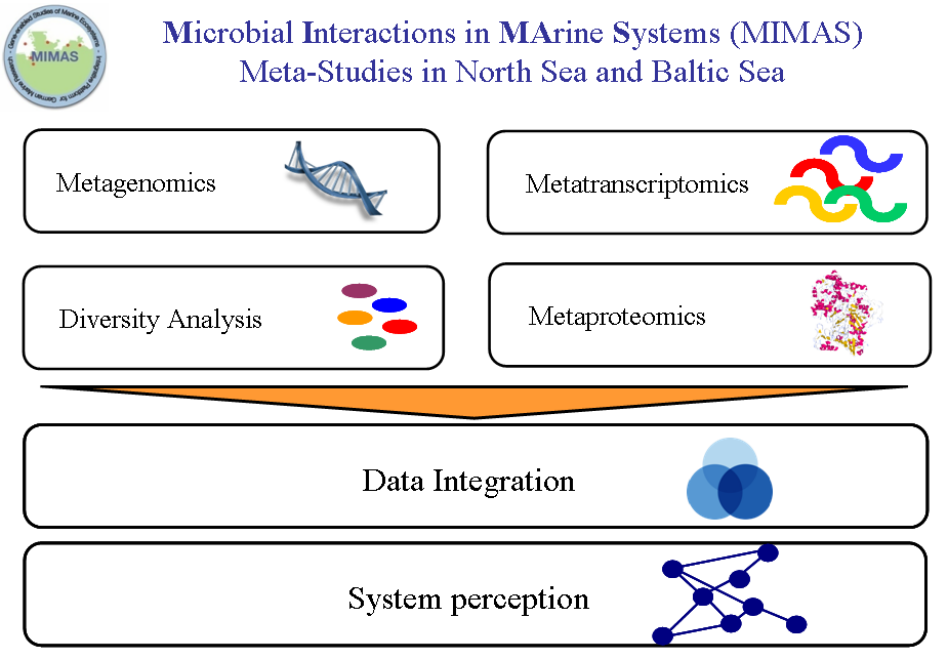


Figure 3: Overview of the sub projects within the MIMAS project.

1.6. Helgoland Roads: one sampling site of choice within the MIMAS project

In the German Bight, about 60 km off the estuaries of the Elbe and Weser rivers, lies the small rocky island Helgoland. Due to its highly diverse marine life and many different easily accessible intertidal and subtidal habitats, this location has attracted many scientists for more than 150 years. In 1873, the Helgoland Roads data series started with daily measurements of temperature and salinity (127,128) which was extended with respect to microbiology parameters in 1963. This long-term series of daily measurements and water sampling is under the supervision of the Biologische Anstalt Helgoland (BAH) and takes place at the 'Kabeltonne' on Helgoland Roads (54°11.3'N, 07°54.0'E). The sampling site is a fixed mooring directly offshore the island, located between the main island Helgoland and the 'Düne' (*engl.* Dune). The name results from, when a buoy was anchored at this particular spot in the past to hold a cable connecting the main island with the Dune.

Nowadays, the Helgoland Roads data series includes physico-chemical parameters (temperatures, salinity, Secchi depth and concentration of dissolved inorganic nutrients: phosphate, nitrate nitrite, ammonium, silicate), as well as biological parameters (qualitative and quantitative data on phytoplankton and microorganisms) (127). It has now been running

continuously for over five decades and has resulted in one of the most important marine data sets in the world with respect to its length and consistency. It is further unique in terms of sampling frequency and number of parameters measured. The Helgoland Roads data series provides scientists with excellent material to monitor food web interaction and analyse the diversity of microbial communities. Further, ecological questions in the course of 'global warming' and other related topics may be answered. For example, the data set already revealed that the average water temperature has risen by 1.13°C since 1962. Likewise a salinity rise of 1.0 PSU has been detected (127). With its fundamental knowledge and on-site equipment, Helgoland Roads stands out as an excellent LTER and provides a promising basis for the characterization of a bacterioplankton community within the MIMAS project.

1.7. *Rhodopirellula baltica* SH1^T: one model organism within the MIMAS project

Pirellula sp. Strain 1, now validly described as *Rhodopirellula baltica* SH1^T (129), was isolated from the water column in the Fjord of Kiel (130). It can be described as a marine aerobic, heterotrophic representative of the environmentally important bacterial order *Planctomycetes*. Members of the latter are abundant in microbial communities in the marine water column and found to be associated with marine snow (131). This leads the assumption of *Planctomycetes* as being key players in carbohydrate metabolism in marine systems (132).
Planctomycetes share several morphological characteristics such as peptidoglycanfree proteinaceous cell walls (133,134), intracellular compartmentalization (135) and a mode of reproduction via budding (136). It is interesting to note that adult *R. baltica* SH1^T exhibit a holdfast substance of so-far unknown chemical composition. The latter can often be observed in natural environments where *R. baltica* SH1^T occurs aggregated and attached to several surfaces such as marine snow (131).
In 2003, the whole genome of *R. baltica* SH1^T was fully sequenced and published (6). With 7,145 Mb and 7,325 open reading frames (ORF) *R. baltica* SH1^T features one of the largest circular bacterial genomes sequenced at that time. The genome annotation revealed genes for the degradation of diverse sugar monomers and complex polysaccharides (e.g. chondroitin sulphate) (129) as well as genes encoding for proteins degradation of the C1-component. What was unexpected, among other things, was the presence of 110 genes encoding for potential sulphatases (6).

1.8. Research aims and thesis structure

RNA based research is a very complex and manifold field due to the different cellular functions of the molecule itself. 16S rDNA biodiversity studies and metatranscriptomics are just two applications among many others indicating the diverse utilization. However, each research field struggles with its own limitations and requires improvements on various levels as elaborated in chapter one. Although the research aims of this thesis are addressing two different RNA based applications, they have a common ancestor: the MIMAS project. With the establishment and improvements of techniques, the analysis and monitoring of bacterial mediated ecosystem functions becomes reliable. The research aims of this thesis address four different aspects as briefly described in chapter 1.9-1.12. Figure 4 provides a schematic overview to guide the reader through the thesis.

1.9. Evaluation of 16S rDNA primer and primer pairs

One of the most crucial steps of PCR amplified 16S rDNA studies is the appropriate selection of the primer pair. Several primers, which are still in use today, have been developed and reviewed many years ago based on the reference sequences available at that time. Nowadays, the amount of sequences in the databases is immense, revealing several new taxonomic groups. In spite of greater diversity, many primers have not been cross-checked and are still widely in use, most likely to ensure comparable studies. Therefore, one research aim of this thesis was to evaluate common universal 16S rDNA primer and primer pairs *in silico* with respect to overall coverage and phylum spectrum for *Bacteria* and *Archaea* (chapter 2, publication 1). Because NGS technology is strongly on the rise to become a standard tool, primer pairs were arranged into groups according to suitable amplicon length addressing the average read length of different sequencing platforms. The gained results were provided to the scientific community in order to serve as a guideline for finding the most suitable primer pair for 16S rDNA analysis in any habitat and for individual research questions.

1.10. NGS based 16S rDNA analysis – proof of concept

Based on the results from the primer evaluation, two primer pairs for 16S rDNA analysis using Roche's 454 pyrosequencing (16S rDNA pyrotags) were chosen for empirical evaluation at the LTER Helgoland Roads (chapter 2, publication 1). The most promising

combination has been applied within a combined study of MIMAS project TP1 and TP3 (chapter 3, publication 2). Initially, 16S rDNA sequences derived from the metagenome confirmed the result from the previous evaluation, and finally, 16S rDNA pyrotag analysis complemented biodiversity information by providing an enhanced resolution up to the group or genus level.

1.11. Functional analysis the of bacterial community at Helgoland Roads in the North Sea

A combined field study of MIMAS project TP1 and TP3 (chapter 3, publication 2) aimed at characterizing a bacterioplankton community by applying a multi 'omic' study. This approach was complemented by metatranscriptomics (chapter 4, publication 3). The major aim of this part included gene expression analysis of the bacterial community at the LTER Helgoland Roads and complements the results of the outcome of the metagenome and metaproteome analysis. Metatranscriptomics confirmed expression profiles with a high confidence, and provided comprehensive data with higher resolution power and taxonomic accuracy. Thereby, key players of the microbial community were easily identified, and in combination with the functional studies, predictive models for bacterioplankton bloom dynamics could be revealed.

1.12. Impact from pure culture studies

In the field of metatranscriptomic research, no standardized protocol is available. Therefore, the first task included the successful set-up of an experimental pipeline. Method development started using *R. baltica* $SH1^T$ (see chapter 1.7) pure cultures as a model system to establish and optimize the different steps. Finally, the gained knowledge was transferred for metatranscriptomic analysis and adapted to environmental samples collected within the MIMAS TP2 project (chapter 4, publication 3). The gained knowledge also influenced the design for a follow-up case study (chapter 5, publication 4). Combined *in vivo* and *in silico* techniques provided first insights into the ecophysiology of planktomycetal sulfatases, which may reflect ecological niches.

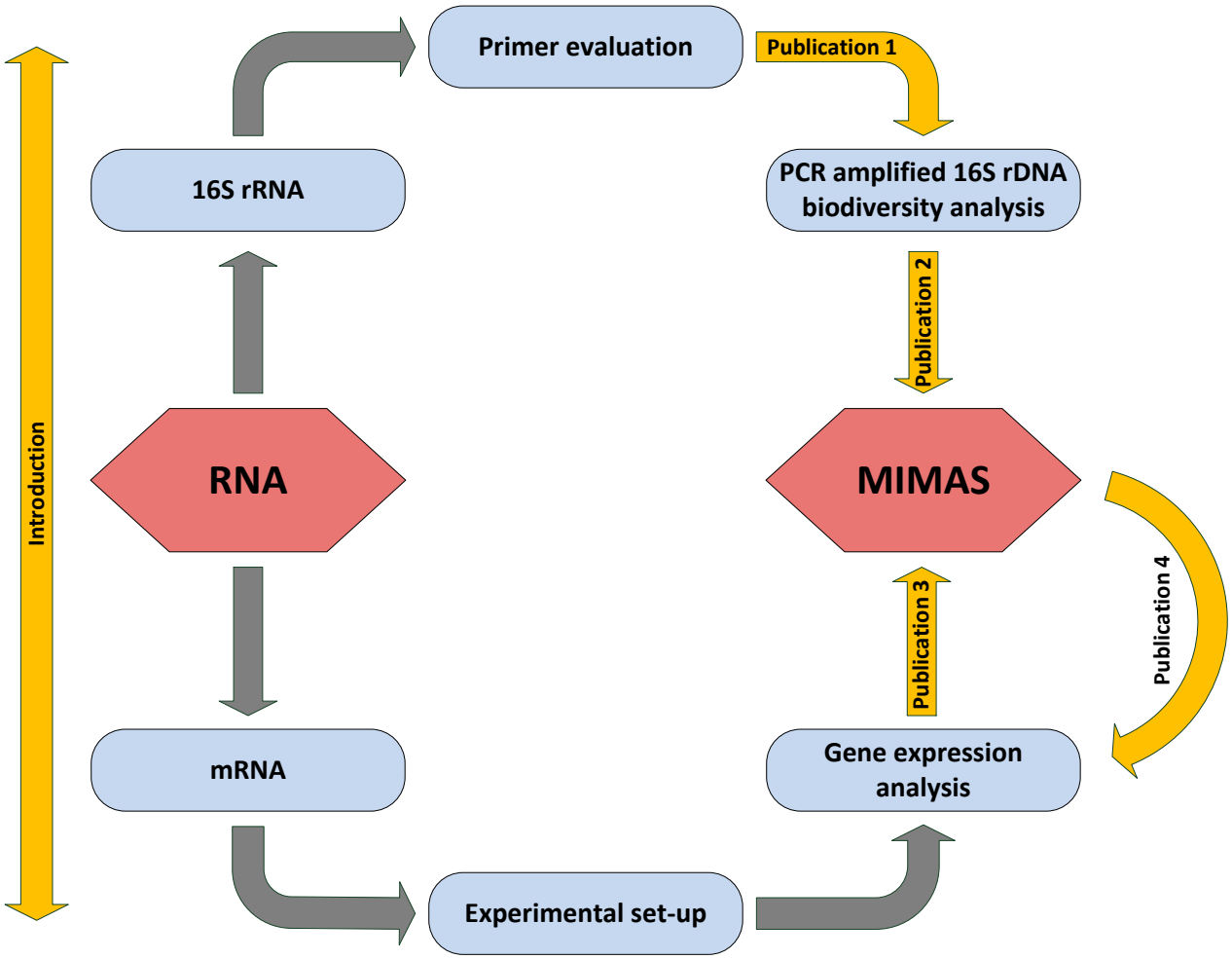


Figure 4: Schematic overview of the thesis structure.

1.13. Publication overview

Chapter 2 – publication 1

Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next generation sequencing based diversity studies

Authors: Anna Klindworth, Elmar Pruesse, Jörg Peplies, Christian Quast, Matthias Horn and Frank Oliver Glöckner

Status: published online on 18th August 2012 in Nucleic Acids Research

Contribution: design of evaluation, performance of laboratory experiments, analysis of data and writing the manuscript

Chapter 3 – publication 2

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

Authors: Hanno Teeling, Bernhard M. Fuchs, Dörte Becher, Christine Klockow, Antje Gardebrecht, Christin M. Bennke, Mariette Kassabgy, Sixing Huang, Alexander J. Mann, Jost Waldmann, Marc Weber, Anna Klindworth, Andreas Otto, Jana Lange, Jörg Bernhardt, Christine Reinsch, Michael Hecker, Jörg Peplies, Frank D. Bockelmann, Ulrich Callies, Gunnar Gerdts, Antje Wichels, Karen H. Wiltshire, Frank Oliver Glöckner, Thomas Schweder, Rudolf Amann

Status: published in Science. 2012; 336(6081):608-611

Contribution: involved in the sampling procedure, based on the previous evaluation (publication 1) design and performance of 16S pyrotag analysis

Chapter 4 – publication 3

Complementary metatranscriptomic analysis of a bacterioplankton bloom in the North Sea

Autors: Anna Klindworth, Alexander Mann, Sixing Huang, Christine Klockow, Jörg Peplies, Christian Quast, Jost Waldmann, Hanno Teeling, Frank Oliver Glöckner

Status: draft to be submitted to Marine Genomics

Contribution: design and performance of laboratory experiments, analysis of data and writing the manuscript

Chapter 5 – publication 4

Expression of sulfatases in *Rhodopirellula baltica* and the diversity of sulfatases in the genus *Rhodopirellula*

Authors: Carl-Eric Wegner, Tim Richter-Heitmann, Anna Klindworth, Christine Klockow, Michael Richter, Tilman Achstetter, Frank Oliver Glöckner and Jens Harder

Status: draft to be submitted to Marine Genomics

Contribution: design, supervision and critical discussion of the experimental part of project, and optimization of experimental procedure in terms of RNA extraction and cDNA synthesis

Other publications not part of this thesis;

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

Authors: Marc Weber, Hanno Teeling, Sixing Huang, Jost Waldmann, Mariette Kassabgy, Bernhard M Fuchs, Anna Klindworth, Christine Klockow, Antje Wichels, Gunnar Gerdts, Rudolf Amann, and Frank Oliver Glöckner

Status: published in ISME Journal 2011; 5(5):918-28

Contribution: Involved in laboratory work and sampling

2. Chapter

Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next generation sequencing based diversity studies

Authors: Anna Klindworth, Elmar Pruesse, Jörg Peplies, Christian Quast, Matthias Horn and Frank Oliver Glöckner

Status: published online on 18th August 2012 in Nucleic Acids Research

Contribution: design of evaluation project, performance of laboratory experiments, analysis of data and writing the manuscript

2.1. Abstract

16S ribosomal RNA gene (rDNA) amplicon analysis remains the standard approach for the cultivation independent investigation of microbial diversity. The accuracy of these analyses depends strongly on the choice of primers. The overall coverage and phylum spectrum of 175 primers and 512 primer pairs were evaluated *in silico* with respect to the SILVA 16S/18S rDNA non-redundant reference dataset (SSURef 108 NR). Based on this evaluation a selection of 'best available' primer pairs for *Bacteria* and *Archaea* for three amplicon size classes (100-400 bp, 400-1000 bp, ≥1000 bp) is provided. The most promising bacterial primer pair (S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21), with an amplicon size of 464 bp, was experimentally evaluated by comparing the taxonomic distribution of the 16S rDNA amplicons with 16S rDNA fragments from directly sequenced metagenomes. The results may be used as a guideline for selecting primer pairs with the best overall coverage and phylum spectrum for specific applications, therefore reducing the bias in PCR based microbial diversity studies.

2.2. Introduction

Understanding microbial diversity has been the ambition of scientists for decades. Because diversity analysis by cultivation is problematic for a significant fraction of *Bacteria* and *Archaea*, culture-independent surveys have been developed. In the past, the most commonly used approach was cloning and sequencing of the 16S ribosomal RNA gene (rDNA) using conserved broad-range PCR primers (32). With the advent of massive parallel sequencing technologies, direct sequencing of PCR amplicons became feasible (46,63,72). In 2006, Roche's 454 GS 20 pyrosequencing (43) became the first high throughput sequencing technology to be successfully applied for large scale biodiversity analysis and was key to uncovering the 'rare biosphere' (22). The continuous development of the technology, offering read lengths of up to 1000 bp nowadays, further improved throughput and resolution of 16S rDNA sequencing (49). Since then, additional high throughput sequencing technologies have become commercially available. The attractiveness of Illumina (44) lies in the reduced per base costs and comparatively high sequencing depth (51), despite having short read lengths. While the major advantage of Ion Torrent (48) are its relatively low cost and rapid sequencing speed. Furthermore, Pacific Bioscience (PacBio) now employs the 'single-molecule real-time' (SMRT) sequencing technology, designed to achieve average read lengths of more than 3,000 bp (62). For a detailed review of sequencing technologies please refer to Loman et al. (124). There is no doubt that the rapid development of sequencing technologies has opened a new dimension in biodiversity analysis, but they also add complexity to the experimental design. The outlined technological differences need to be carefully considered when analysing the results to approximate the 'natural' diversity distribution.

The most critical step for accurate rDNA amplicon analysis, however, is the choice of primers (63,137). Using suboptimal primers, or more precisely primer pairs, can lead to under-representation (69) or selection against single species or even whole groups (24,42,70). Using inappropriate primers consequently leads to questionable biological conclusions (24,71,138).

In this study, 175 broad range 16S rDNA primers and 512 primer pairs were investigated *in silico* with respect to overall coverage and phylum spectrum for *Bacteria* and *Archaea* as well as amplicon length. Primer sequences were compared to all 376,437 16S/18S rDNA sequences available in the SILVA non-redundant reference database (SSURef NR) release 108 (36). For consistency, all primers were renamed according to the primer nomenclature suggested by Alm et al. (139). Two pairs of bacterial PCR primers were selected for empirical evaluation at the field station Helgoland Roads (North Sea). Finally, the obtained results were compared with diversity estimates from previous metagenome studies (140).

2.3. Material and Methods

Primer Nomenclature

Primers were re-named according to Alm et al. (139). Each name is composed of seven dash-separated parts, describing: the target gene, the rank of the target group, the target group, the target position within the gene, the primer version, the target strand and the length of the primer. For illustration, the seven parts comprising the primer name ‘S-D-Bact-0338-a-A-18’ are to be interpreted as follows:

1) An indication of the target gene. In this case, ‘S’ for small subunit rDNA (S).
2) An indication of the largest taxonomic group targeted by the PCR primer. For example, ‘D’ for domain level.
3) An abbreviated description, limited to three to five letters, of the specific taxonomic or phylogenetic group targeted by the primer. For example, ‘Bact’ for the domain *Bacteria*.
4) A four-digit number indicating the 5’position of the sense strand. For example ‘0338’ stands for start position 338 in the *Escherichia coli* system of nomenclature (141).
5) A single lowercase letter indicating the version of the probe. For example ‘a’ for a first version.
6) A single uppercase letter indicating whether the probe sequence is identical to the DNA sense strand (S) or to the antisense (A) strand
7) A number indicating the length of the PCR primer, e. g. 18 bases in the example.

Nomenclature for in silico evaluation

In this study, the term ‘coverage’ refers to the percentage of matches for a given taxonomic path. Taxonomic paths were considered ‘not covered’ if their coverage was below 50%. The term ‘phylum spectrum’ refers to the number of matched phyla. For example, if a primer or primer pair covers the majority of all phyla it is described as having a ‘large phylum spectrum’.

Selection of primers

A total of 175 forward and reverse 16S rDNA primers were chosen for the *in silico* evaluation. Primer sequences were either obtained from a literature survey or provided by the SILVA user community in response to a poll on the ARB/SILVA mailing list in January 2012 (Supplementary Table 1). Only primers with an overall coverage above 75% for either *Bacteria* or *Archaea* were considered for primer pair analysis. All primers are available in probeBase, a comprehensive online database for rRNA-targeted oligonucleotides, at http://www.microbial-ecology.net/probebase/ (24).

Selection criteria for primer pairs

Primer pairs were chosen according to annealing temperatures, overall coverage of variable regions and amplicon length. Annealing temperatures were calculated with OligoCalc (142). Primer pair combinations with annealing temperature differences of less than 5°C were accepted as pairs. Suitable primer pairs were organized into three different groups (Supplementary Table 8): *Group Short* (*Group S*) generates 100-400 bp fragments. *Group Middle* (*Group M*) generates 400-1000 bp fragments. *Group Long* (*Group L*) generates fragments ≥1000 bp. A total of 512 primer combinations were evaluated. The best 30 bacterial primer pairs in each group and all archaeal primer pairs with a combined overall coverage >70% were analyzed in detail.

In silico evaluation of primers and primer pairs

Primer evaluation was based on two datasets: Firstly, the non-redundant SILVA Reference database (release SSURef 108 NR) containing 376,437 sequences. The SILVA SSURef 108 NR was prepared from all SSU sequences longer than 1,200 bp for *Bacteria* and *Eukaryota* and longer than 900 bp for *Archaea.* Sequences are required to have a SINA (143) alignment quality value better than 50 (36). Redundant sequences were removed by clustering with UCLUST (144) using a 99% identity criterion. A second SSU database was prepared from the Global Ocean Survey (GOS) (145,146) metagenomes using the SILVA pipeline. Alignment was attempted with SINA for all GOS reads and all sequences with an alignment quality of at least 30 and a minimum length of 300 were retained, yielding a dataset of 10,945 sequences. Taxonomic classifications for each read were applied as described below.

Primer matching was executed using the probe match function of the ARB PT server (147) at two levels of stringency, allowing zero or one mismatch, respectively. For each primer and

stringency level the database entries were separated into three groups: 1) matches, 2) mismatches and 3) unknown. The match status was considered to be unknown if no sequence data was available at the match position of the respective primer. Furthermore, only sequences corresponding to the primer at the intended position where considered to be matches. From these numbers, coverage was computed as the matched fraction of entries either matches or mismatches, excluding entries for which the match status was unknown. Individual coverages were computed for all taxa. When computing the combined coverage of forward and reverse primer pairs, an entry was considered to have unknown match status if the match status for either of the two primers was unknown. Likewise, the pair was only considered to be a match if both primers matched at the intended match position.

Detailed information for each analyzed primer and primer pair are provided in the supplementary material online (single primer: Supplementary Table 2-7; primer pairs: Supplementary Table 9-38). All scripts and SQL queries as well as database dumps are available online at http://www.arb-silva.de/download/archive/primer_evaluation.

Sample site and collection of water samples

Sample collection was carried out as part of the 'multi omic' approach of the MIMAS (Microbial Interaction in MArine Systems) project (www.mimas-project.de). Surface water was collected on 11th February 2009 and weekly from 31th of March 2009 until October 2009. Water samples (total volume 360 l) from the Kabeltonne site at Helgoland Roads in the North Sea (54°11.18′N, 7°54.00′E) were collected at a depth of 0.5 m and processed immediately at the Biological Station Helgoland. The water was pre-filtered through a 10 µm and a 3 µm pore-size filter. For harvesting a 0.2-µm-pore-size filter was used. At each time point 10 l and 15 l of seawater were filtered onto eight filters for genomic DNA extraction. All filters were stored at -80°C until future usage. Details can be found in Teeling et al. (140). In this study, 16S rDNA pyrotag analysis with Roche's 454 FLX Titanium technology was performed using samples from: 11.02.2009, 07.04.2009 and 14.04.2009. Results from 16S rDNA diversity analysis gained from metagenome studies of the same sampling dates (140) were used for comparison.

DNA extraction

Genomic DNA was directly extracted from filters as described in Zhou et al. (148) with the following modifications: all extraction steps were performed with 50 µl proteinase K (10 mg/ml), and after isopropanol precipitation, pelleted nucleic acids were obtained by

centrifugation at 50,000 g for 30 min at room temperature. The genomic DNA was stored at -20°C until PCR amplification and metagenomic sequencing were carried out.

Amplification

Per sample, two separate PCR reactions were performed in order to test two bacterial primer pairs for 16S rDNA amplification. Primer pairs were: (A): S-D-Bact-0341-b-S-17, 5'-CCTACGGGNGGCWGCAG-3'(149), and S-D-Bact-0785-a-A-21, 5'-GACTACHVGGGTATCTAATCC-3 (149), and (B): S-D-Bact-0008-a-S-16, 5'-AGAGTTTGATCMTGGC-3'(150), and S-D-Bact-0907-a-A-20, 5'-CCGTCAATTCMTTTGAGTTT-3' (151). The reaction was carried out in 50 µl volumes containing 0.3 mg/ml BSA (Bovine Serum Albumin), 250 µM dTNPs, 0.5 µM of each primer, 0.02 U Phusion High-Fidelity DNA Polymerase (Finnzymes OY, Espoo, Finland) and 5x Phusion HF Buffer containing 1.5 mM $MgCl_2$. The following PCR conditions were used: initial denaturation at 95°C for 5 min, followed by 25 cycles consisting of denaturation (95°C for 40 sec), annealing (2 min) and extension (72°C for 1 min) and a final extension step at 72°C for 7 min. Annealing temperature for primer pair (A) was set at 55°C and for (B) at 44°C. PCR products were purified with a QiaQuick PCR purification kit (QIAGEN, Hilden, Germany). The quantity and quality of the extracted DNA were analyzed by spectrophotometry using an ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, DE) and by agarose gel electrophoresis. The PCR products were stored at -20°C for sequencing.

Sequencing

The pyrosequencing reactions were performed at LGC Genomics GmbH, Berlin, Germany. All sequencing reactions were based upon FLX –Titanium chemistry (Roche/454 Life Sciences, Branford, CT 06405, USA; www.454.com) and all methods were performed using the manufacturers' protocol. Briefly, genomic DNA from metagenome studies (140) as well as PCR amplified DNA fragments were checked for quality on a 2% agarose gel. 500 ng of each sample was then used for the sequencing library. In a minor modification to the protocol, no size selection of the fragments was performed. The fragments were subjected to end repair and polishing. An extra A was added to the ends of the fragments and the Roche Rapid Library adaptors were ligated on to the fragments as described in the Roche Rapid Library Preparation Manual for GS FLX Titanium Series, October 2009, Rev. Jan. 2010 (Roche/454 Life Sciences, Branford, CT 06405, USA; www.454.com). After subsequent emulsion PCR

the fragment libraries were processed and sequenced according to the Roche protocols. The resulting sequences were processed using the standard Roche software for base calling, trimming of adaptors and quality trimming (Genome Sequencer FLX System Software Manual version 2.3, Roche/454 Life Sciences, Branford, CT 06405, USA; www.454.com). For PCR amplified DNA fragments, per sample two distinct PCR reactions were sequenced on 1/8 of picotiter plate (PTP). Raw data was stored as FNA file. Sequences were submitted to INSDC (EMBL-EBI/ENA, Genbank, DDBJ) with accession number ERP001031. For metagenomics two full PTPs per sample were sequenced. Metagenome sequences were published by the MIMAS project (140) and can be obtained from INSDC with accession number ERP001227.

Identification and taxonomic classification of 16S rDNA fragments

Unassembled sequence reads from both SSU rRNA gene PCR amplicons (pyrotags) and metagenome sequencing were preprocessed (quality control and alignment) by the bioinformatics pipeline of the SILVA project (36). Briefly, reads shorter than 200 nucleotides or with more than 2% of ambiguities or more than 2% of homopolymers were removed. Remaining reads from amplicons and metagenomes were aligned against the SSU rDNA seed of the SILVA database release 108 (http://www.arb-silva.de/documentation/background/release-108/) (36) using SINA (143). Unaligned reads were not considered in downstream analysis to eliminate non 16S rDNA sequences.

Remaining PCR amplicons were separated based on the presence of aligned nucleotides at *E. coli* positions of the respective primer binding sites instead of searching for the primer sequences itself. This strategy is robust against sequencing errors within the primer signatures or incomplete primer signatures. This separation strategy works because the amplicon size of one primer pair is significant longer, with overhangs on both 3' and 5' site, compared to the amplicon of the second primer pair. With this approach the need for barcoding during combined sequencing of 16S pyrotags derived from different PCR reactions on the same PTP lane was avoided. FASTA files for each primer pair of the separated samples are available online http://www.arb-silva.de/download/archive/primer_evaluation.

Reads of the filtered and separated 16S pyrotag datasets as well as metagenomes were dereplicated, clustered and classified on a sample by sample basis. Dereplication (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 0.98 (wordsize 8). The longest read of each cluster was used as a reference for taxonomic classification, which was done using a local BLAST search against the SILVA SSURef 108 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 default settings. The full SILVA taxonomic path of the best BLAST hit was assigned to the reads if the value for (% sequence identity + % alignment coverage)/2 was at least 93. 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 (as identified in the previous analysis step) to finally obtain (semi-) quantitative information (number of individual reads representing a taxonomic path). Raw output data are available in the supplementary material in Supplementary Tables 48-50.

Adjustment of the total number of sequences reads to smaller subset by random re-sampling

Sequencing depth may infringe on the comparability of the resulting taxonomic resolution. To verify that the results derived from the 16S pyrotags were not an artefact of deep sequencing, the total number of 16S pyrotags was reduced until roughly equal amounts of classified pyrotags and classified metagenome reads remained for each sample. Three subsets of each 16S pyrotag sample were adjusted by withdrawing equal amounts of sequences randomly without replacement. Raw output data are available in the supplementary material online (Supplementary Table 51-52). An analogue approach was described in Gilbert et al. (152).

2.4. Results and Diskussion

In silico evaluation of 16S rDNA primers

The overall coverage of 175 single primers was evaluated for all three domains of life (Supplementary Table 1). Additionally for *Bacteria* and *Archaea* the phylum spectrum was investigated with respect to zero and one mismatch (Supplementary Table 2-5). *Eukaryota* are only considered on domain level (Supplementary Table 5-6). 122 single primers passed the 50% overall coverage threshold with 31, 51 and 1 primer(s) specific for the domain *Archaea* (A), *Bacteria* (B) and *Eukaryota* (E), respectively. At one–mismatch-stringency the total number increased to 150 eligible primers.

For *Archaea*, primers S-D-Arch-0519-a-A-19 (A: 91.3%, B: 0.1%, E: 1.0%) and S-D-Arch-0787-a-A-20 (A: 87.4%, B: 7.8%, E: 0.0%) stand out. This is in line with a recent study by

Wang et al. (70). The highest overall coverage and specificity for the domain *Bacteria* was detected for the primers S-D-Bact-1061-a-A-17 (A: 2.9%, B: 96.4%, E: 0.0%) and S-D-Bact-0564-a-S-15 (A: 16.3%, B: 96.0%, E: 0.0%). Furthermore, 39 primers show relatively high overall coverage for more than one domain. For instance, S-*-Univ-0515-a-S-19 (A: 54.5%, B: 95.4%, E: 92.2%) detects all three domains and S-D-Bact-0787-b-A-20 (A: 89.9%; B: 90.6%; E: 0.0%) targets *Bacteria* and *Archaea* as recently reported (153).

It has previously been asserted (70) that the primers S-*-Univ-0789-a-S-18 (A: 86.1%, B: 6.8%, E: 0.0%) and S-*-Univ-0906-a-S-17 (A: 83.7%, B: 0.3%, E: 76.8%) target *Bacteria* and *Archaea*. Contrary to this, with only 6.8% and 0.3% overall coverage of the domain *Bacteria*, but 86.1% and 83.7% overall coverage of the domain *Archaea*, respectively, our results confirm the original intention of both primers to be specific for the domain *Archaea* (154,155). However, if one mismatch is tolerated, S-*-Univ-0789-a-S-18 (A: 96.0%, B: 93.0%, E: 0.0%) targets *Archaea* and *Bacteria*. S-*-Univ-0906-a-S-17 (A: 93.2%, B: 49.8%, E: 0.0%) still fails to pass our 50% threshold.

The primer sequence of S-*-Univ-0779-a-S-20 (A: 0.0%; B: 0.0%, E: 0.0%) is misspelled in Wang et. al. (70). Allowing one mismatch increases the overall coverage to A: 64.8%, B: 6.8%, E: 77.6% and indicates that the correct primer sequence targets *Archaea* and *Eukaryota*.

A direct comparison of our results with the studies of Huws et al. (156) and Baker et al. (69) is not possible, as the overall coverage of the primers is not given. Nossa et al. (32) restricted their evaluation to a single habitat. Walter et al. (153) analysed a total of only four primers.

In respect to detailed phylum coverage (see Supplementary Table 2-5) it should be noted that the numbers of sequences present in a phylum affects the values for phylum coverage. If the majority of a small phylum (e.g. *Caldiserica* with 61 sequences) is targeted, the coverage will be higher than for a member rich phylum (e.g. *Firmicutes* with 84,910 sequences). Similar effects occur for phyla in which only a small number of sequences contain sequence information at the primer position of interest.

In silico evaluation of primer pairs

When combining forward and reverse primers, the bias of single primers can accumulate. To minimize the overall bias, primers with similar overall coverage and phylum spectrum must be used. Using the 75% overall coverage criterion, 86 single primers qualify for primer pair analysis. In order to get suitable combinations for the different sequencing technologies, primer pairs were organized into three groups according to their amplicon length (Supplementary Table 8). *Group S(mall)* could be of particular interest for Illumina (44) and

Ion Torrent (48) sequencing. Primer pairs of *Group M(iddle)* are suitable for Roche's 454 (157) technology. *Group L(arge)* primer pairs are useful for sequencing methods such as PacBio (62) as well as for creating classical clone libraries. A total of 512 primer combinations were evaluated. Again, the focus of this evaluation was *Archaea* and *Bacteria*. *Eukaryota* are only considered on domain level.

Assuming that a standard PCR can tolerate up to two mismatches between the primer and its target (32), results with one mismatch are also taken into account. However, it should be noted that a primer mismatch can result in a biased picture of the bacterial diversity (158) and preferential amplification might lead to under-representation of important members of a community (69,158).

In silico evaluatation of primer pairs suitable for Illumina and Ion Torrent sequencing (*Group S*)

Only 12 archaeal primer pairs have an overall coverage above 70%. The best results with an overall coverage of 76.8% are obtained with S-D-Arch-0349-a-S-17/S-D-Arch-0519-a-A-16 (A: 76.8%, B: 0.0%, E: 0.0%) (Supplementary Table 9). This pair generates an amplicon length of 185 bp which spans the hypervariable (HV) region three. The evaluation revealed that it misses five out of eight phyla: Ancient Archaeal Group (AAG), GoC-Arc-109-D0-C1-M0, *Korarchaeota*, Marine Hydrothermal Vent Group 2 (MHVG-2) and *Nanoarchaeota*. The remaining three archaeal phyla are detected (*Crenarchaeota*, Marine Hydrothermal Vent Group 1 (MHVG-1) and *Euryarchaeota*). With one mismatch allowed, overall coverage for *Archaea* increases to A: 91.0%, B: 0.0%, E: 0.1% now covering additionally *Korarchaeota* and MHVG-2 (Supplementary Table 10). However, in the case of *Korarchaeota* detailed analysis of the primer target position revealed a 3'end mismatch of the forward primer, which is known to affect amplification. *Nanoarchaeota* and AAG show three mismatches. Moreover, PCR has to tolerate up to four mismatches of the forward primer to amplify GoC-Arc-109-D0-C1-M0. In summary, S-D-Arch-0349-a-S-17/S-D-Arch-0519-a-A-16 generates short amplicons, has a comparatively high overall coverage by detecting up to four out of eight archaeal phyla and excellent domain specificity. Hence, this primer pair shows the most promising results for Illumina and Ion Torrent sequencing.

For *Bacteria*, the primer pair S-D-Bact-0341-b-S-17/S-D-Bact-0515-a-A-19 (A: 0.0%, B: 91.2%, E: 0.0%) has the highest overall coverage (Supplementary Table 11). Detailed analysis reveals that 10 phyla are not detected (*Armatimonadetes, Chlamydiae, Caldiserica,* Hyd24-12, GOUTA4, Kazan-3B-28, SM2F11, as well as Candidate divisions WS6, OP11,

TM7 and OD1). If one mismatch is tolerated some *Archaea* (A: 44.6%, B: 96.7%, E: 0.2%) as well as seven additional phyla are detected (Supplementary Table 12), but amplification of Candidate divisions OP11 and WS6 as well as *Armatimonadetes* remains unlikely. In all three cases, the mismatch position of the forward primer is located at the 3'end. For Candidate divisions OP11 and WS6, the reverse primer would need to tolerate three mismatches. These findings are in line with the conclusions of Baker et al. (159), who claim that no domain specific primer exists or can be designed that matches all bacterial 16S rDNA sequences.

The best candidate for the domain *Bacteria* is S-D-Bact-0564-a-S-15/S-D-Bact-0785-b-A-18. This primer pair has a slightly lower overall coverage for *Bacteria* (A: 14.6%, B: 89.0%, E: 0.0%) compared to the previous candidate but only fails to detect four bacterial phyla (*Chloroflexi, Elusimicrobia,* BHI80-139 and Candidate division OP11). With one allowed mismatch (A: 57.1%, B: 95.2%, E: 0.0%), only Candidate division OP11 sequences remain undetected due to a 3'end mismatch of both primers. Please note that one mismatch may also lead to amplification of archaeal 16S rDNA sequences. Based on the promising phylum spectrum we are in favor of this primer pair in comparison to the previous described S-D-Bact-0341-b-S-17/S-D-Bact-0515-a-A-19. In summary, S-D-Bact-0564-a-S-15/S-D-Bact-0785-b-A-18 generates an amplicon of 253 bp covering the fourth HV region and satisfies with a high overall coverage and reasonably good domain specificity. Hence, it is recommended for *Bacteria.*

Two primer pairs target the domains *Bacteria* and *Archaea*: S-D-Arch-0519-a-S-15/S-D-Bact-0785-b-A-18 (A: 88%; B: 89.1%, E: 0.7%) and S-D-Arch-0519-a-S-15/S-D-Bact-0785-a-A-21 (A: 86.5%; B: 87.1%, E: 0.0%). Within the bacterial domain, those two primer pairs cover 49 out of 59 phyla. The coverage for *Chlamydiae, Caldiserica, Chloroflexi*, SM2F11, Kazan-3B-28, BHI80-139 and Candidate divisions WS6, OP11, TM7 and OD1 is below 50%. If one mismatch is tolerated, seven additional phyla are detected and overall coverage increases for S-D-Arch-0519-a-S-15/S-D-Bact-0785-b-A-18 (A: 94.9%; B: 95.1%, E: 1.6%) and S-D-Arch-0519-a-S-15/S-D-Bact-0785-a-A-21 (A: 94.6%, B: 94.8%, E: 0.7%). Amplification of Candidate divisions WS6, TM7 and OP11 remains unlikely. The mismatch position of S-D-Arch-0519-a-S-15 is located at the 3'end in case of Candidate divisions WS6 and TM7. For Candidate division OP11, both reverse primers show a 3'end mismatch. For *Archaea*, each primer pair fails to detect four out of eight phyla (AAA, MHVG-1 and MHVG-2 and *Nanoarchaeota*), which is reduced to one (*Nanoarchaeota*) if one mismatch is allowed. The continuous failure of primers to detect *Nanoachaeota* is not surprising, due to the majority of *Archaea* specific primers being designed prior to the discovery of the *Nanoarchaeota* (69).

Detailed analysis of the mismatch positions reveals one internal mismatch for AAA, MHVG-1 and MHVG-2 but three mismatches for *Nanoarchaeota*. Addition of *Nanoarchaeota* specific primers (160) is recommended. Previous evaluation showed S-P-Nano-0008-a-S-16 and S-P-Nano-1390-a-A-17 to be highly specific for *Nanoarchaeota* (Supplementary Table 2). Note that these primers generate almost full length sequences. In summary, both primer pairs can be recommended for amplification. They generate amplicons specific for *Bacteria* and *Archaea* with an average length of 278 bp that spans the HV region four.

In silico evaluatation of primer pairs suitable for sequencing technologies like Roche's 454 (*Group M*)

No archaeal specific primer pair achieves a full phylum spectrum (Supplementary Table 15). S-D-Arch-0519-a-S-15/S-D-Arch-1041-a-A-18 (A: 76.6%, B: 0.0%, E: 0.0%) shows the best results with respect to a relatively high overall coverage coupled with a high domain specificity. This primer pair covers two out of eight phyla (Crenarchaeota and Euryarchaeota), but the phylum spectrum increases remarkably to six detected phyla if one mismatch is allowed (A: 92.8%, B: 0.0%, E: 0.0%). Detection of the four additional phyla (AAG, Korarchaeota, MHVG I and MHVG II) is likely due to a middle mismatch position in the reverse primer. Amplification of GoC-Arc-109-D0-C1-M0 and Nanoarchaeaota remains challenging due to more than one mismatch. In summary, S-D-Arch-0519-a-S-15/S-DArch-1041-a-A-18 is the most suitable primer pair with a 540 bp amplicon spanning HV regions 4-6 and excellent domain specificity. The frequent use of HV region six in diversity analysis makes this pair particularly interestingfor comparative analysis (28,152,161).

For the domain *Bacteria*, several domain specific primer pairs attain high overall coverage, but 27 out of 30 fail to detect more than 10 phyla (Supplementary Table 17). The three best pairs are S-D-Bact-0341-b-S-17/S-D-Bact-1061-a-A-17 (A: 0.0%, B: 91.9%, E: 0.0%), S-D-Bact-0564-a-S-15/S-*Univ-1100-a-A-15 (A: 8.0%, B: 92.7%, E: 0.0%) and S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21 (A: 0.5%, B: 86.2%, E: 0.0%). Although the first two show higher overall coverage, the latter exhibits a larger phylum spectrum. S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21 only fails to detect seven bacterial phyla (Hyd24-12, GOUTA4, *Armatimonadetes, Chloroflexi*, BHI80-139 and Candidate divisions OP11 and WS6). If one mismatch is tolerated (A: 64.6%, B: 94.5%, E: 0.1%), amplification of four additional phyla is likely (*Chloroflexi*, BHI80-139, Hyd24-12 and GOUTA4). However, some archaeal sequences are also detected. Detailed analysis reveals that only the coverage for Candidate division OP11 remains below the 50% threshold (Supplementary Table 18). Besides four

mismatches for the reverse primer, the mismatch positions in both primers are located towards the 3' end. Moreover, amplification of *Armatimonadetes* and Candidate division WS6 is unlikely due to the 3' end mismatch position of the forward primer. Although not covering the complete phylum spectrum, the pair S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21 shows the best combination of domain and phylum coverage and can thus be recommended for 464 bp amplicons covering the HV regions 3-4.

S-D-Bact-0785-a-S-18/S-*-Univ-1392-a-A-15 (A: 72.3%; B: 74.1%, E: 0.0%) qualifies as a suitable primer pair for *Bacteria* and *Archaea*. With no mismatches it only fails to detect *Nanoachaeota* and expands to full archaeal phylum spectrum if one mismatch is tolerated. Detailed analysis revealed that none of the mismatch positions are located towards the 3'end, which should allow amplification. For *Bacteria*, an overall coverage of 76.3% is achieved but this pair fails to detect nine phyla (*Chloroflexi*, SM2F11, HDB-SIOH1705, BD1-5, EM19, BHI80-139, Candidate divisions OP11, SR1, OD1 as well as *Epsilonproteobacteria*). Allowing one mismatch results in an increased overall coverage (A: 79.0%, B: 86.1%, E: 1.3%) and the additional detection of six phyla due to internal mismatches. Only the coverage of HDB-SIOH17005, SM2F11 and Candidate division OP11 remains below the 50% threshold. In summary, with an amplicon length of 608 bp and detection of HV region 5-8 this primer pair qualifies to target *Bacteria* and *Archaea*.

This detailed evaluation also demonstrates that reverse and forward primers with individual high coverage do not automatically qualify as an optimal primer pair. For instance, S-D-Bact-0347-a-S-19 (A: 0.0%, B: 86.1%, E: 0.0%) and S-D-Bact-0785-a-A-19 (A: 8.5%, B: 86.4%, E: 0.0%) have been designed and approved by the Human Microbiome Project for analysing the foregut microbiome (32). Based on promising results within the human habitat, they suggested that this primer pair may be a good candidate to access the bacterial diversity in any habitat (32). However, our evaluation reveals a lower overall coverage of A: 0.0%, B: 76.5%, E: 0.0% and detection of only 25 out of 59 bacterial phyla if they act as a primer pair. Even if one mismatch is allowed (A: 0.0%, B: 90.6%, E: 0.0%) this primer pair still fails to detect 17 phyla (*Armatimonadetes, Chlamydiae, Dictyoglomi, Planctomycetes, Verrucomicrobia, Spirochaetes, Lentisphaerae,* HDB-SIOH1705, LD1-PA38, NPL-UPA2, Hyd24-12 and SM2F11, as well as Candidate divisions OP11, WS6, BRC1, OD1, WS3 and OP3).

In silico evaluation of primer pairs suitable for sequencing technologies such as PacBio SMRT or classical clone libraries (*Group L*)

For fragments >1000 bases we could not find an archaeal primer pair with both an overall coverage of over 70% and a satisfying phylum spectrum (Supplementary Table 21). The majority detects only the two sequence-rich phyla, *Crenarchaeota* and *Euryarchaeota.* S-D-Arch-0349-a-S-17/S-*-Univ-1392-a-A-15 (A: 65.8%, B: 0.0%, E: 0.0%) has the highest overall coverage. Detailed analysis revealed that this pair fails to detect six out of eight phyla (AAG, GoC-Arc-109-D0-C1-M0, *Korarchaeota*, MHVG-1, MHVG-2 and *Nanoarchaeota*) (Supplementary Table 21). Although performance increases slightly when one mismatch is allowed (A: 76.0%, B: 0.0%, E: 0.1%), the coverage for three phyla (AAG, GoC-Arc-109-D0-C1-M0 and *Nanoarchaeota*) remains below 50% (Supplementary Table 22). In addition, a 3' mismatch of the forward primer hampers amplification of *Korarchaeota.* In summary, this primer pair cannot be recommended. Similar results are obtained for the other archaeal primer pairs of *Group L*.

The bacterial primer pairs show more satisfying results (Supplementary Table 23). S-D-Bact-0008-c-S-20/S-D-Bact-1391-a-A-17 (A: 0.1%, B: 78.0%, E: 0.0%) has a high overall coverage and detects 55 out of 59 phyla. The four phyla with below-threshold coverage are *Chlamydiae,* WCHB1-60, Candidate division SR1 and OP11. If one mismatch is allowed, overall coverage increases to A: 0.1%, B: 86.2%, E: 0.0% and Candidate division OP11 is now likely to be detected due to an internal mismatch. S-D-Bact-0008-c-S-20/S-D-Bact-1046-a-A-19 (A: 0.0%, B: 81.3%, E: 0.0%) achieves the highest overall coverage but fails to detect eight phyla (S2R-29, SM2F11, *Chlamydiae, Thermotogae*, WCHB1-60, Kazan-3B-28, EM19, Candidate division OP11 and *Epsilonproteobacteria*). Remarkably, this is mostly compensated if one mismatch is allowed. However, amplification of some sequences belonging to Candidate division OP11 and WHCBI-60 is unlikely due to 3'end mismatches. Moreover, the reverse primer fails to detect SM2F11 due to two mismatches of which one is located towards the 3'end. *Chlamydiae* remains undetected due to three internal mismatches of the forward primer. The promising results and excellent domain specificity of both primer pairs are depreciated by the fact that they only span HV regions 1-6 and 1-8, respectively. Nevertheless, if an amplicon length of <1400 bp is sufficient we are in favour of both primer pairs.

For nearly full length sequences (>1400 bp) we recommend S-D-Bact-0008-a-S-16/S-D-Bact-1492-a-A-16 (A: 0.2%, B: 77.1%, E: 0.0%). This domain specific primer pair spans HV

regions 1-9 and covers 52 out of 59 bacteria phyla. The missing phyla are: GAL08, Kazan-3B-28, *Chlamydiae*, *Dictyoglomi*, WCHB1-60, MVP-21 and *Caldiserica*. One mismatch (A: 0.2%, B: 86.8%, E: 0.0%) allows additional detection of *Caldiserica* and *Dictyoglomi* due to an internal mismatch. The remaining five phyla have either more than two mismatches or, in case of *Chlamydiae*, the forward primer has a 3'end mismatch. In the past, S-D-Bact-0008-a-S-16/S-D-Bact-1492-a-A-16, which is commonly known as GM3/GM4, has been intensively used for clone library based studies from different habitats (33-35). Thus plenty of data for comparative analysis is available. However, the high number of sequences originally obtained with the GM3/GM4 pair is also likely to have artificially inflated the coverage values we obtained. Ideally, sequences obtained with a given primer should be excluded when evaluating that same primer.

In silico re-evaluation of primer pairs using a PCR free metagenome database

The majority of the sequences in specialised 16S/18S rDNA databases such as SILVA (36), greengenes (37) or RDP II (38) are a result of prior PCR amplification. In order to calibrate our previous analysis, re-evaluation of the results using the publicly available Global Ocean Sampling (GOS) database was performed. The initial GOS dataset consisted of 6.3 billion bp of Sanger sequence reads (145) and has recently been augmented by samples from the Atlantic and Indian Oceans (162). Although it is limited to the marine habitat, it is the most comprehensive dataset that provides a reasonable amount of relatively long fragments necessary for primer evaluation.

A total of 10,685 16S/18S rDNA sequences were extracted from the GOS dataset. 95% of the reads range between 900 and 1200 bp in length; the average length was 1053 bp. However, the bacterial fraction was dominant, consisting of 9965 sequences, compared to only 290 archaeal and 439 eukaryotic 16S and 18S sequences, respectively. Thus the results for *Archaea* and *Eukaryota* are uncertain and should only be seen as complementary information. In addition to the limited number of sequences, only a subset of phyla is present. For example, for *Archaea* 288 sequences belong to *Crenarchaeota* (63 sequences) and *Euryarchaeota* (225 sequences). The remaining two sequences could be assigned to AAG and MHVG-1, respectively. For *Korarchaeota*, GoC-Arc-109-D0-C1-M0, MHVG-2 and *Nanoarchaeota*, no sequences are present.

For the domain *Bacteria*, the 9956 reads span 28 out of 59 phyla. The majority belong to *Actinobacteria* (1006 sequences), *Bacteroidetes* (785 sequences), *Cyanobacteria* (805

sequences) and *Proteobacteria* (6655 sequences). Other member rich phyla such as *Firmicutes* (167 sequences) and *Acidobacteria* (29 sequences) are only present in low numbers. The lack of a full phylum spectrum clearly limits the re-evaluation and prevents direct comparisons with our previous results. The much lower and also varying number of sequences in the respective target regions affects the results as well. Furthermore primer pairs of *Group L* had to be excluded from the re-evaluation due to the lack of sufficient numbers of long sequences.

In the previous evaluation for *Group S*, the archaeal primer pair S-D-Arch-0349-a-S-17/S-D-Arch-0519-a-A-16 (A: 76.8%, B 0.0%, E: 0.0%) was proposed as a suitable pair for amplicon sequencing of <400 bases. Re-evaluation based on the GOS dataset again yielded the highest overall coverage (A: 74.5%, B: 0.0%, E: 1.2%) and excellent domain specificity. The recommended bacterial primer pair S-D-Bact-0564-a-S-15/S-D-Bact-0785-b-A-18 (A: 0.0%, B: 83.4%, E: 0.0%) also performs well. Tolerating one mismatch still confirms domain specificity (A: 10.6%, B: 86.2%, E: 0.0%). Unfortunately, detailed comparison on phylum level proved difficult. For example, within the SILVA database, 84,910 *Firmicutes* sequences of sufficient length are present and 91.8% of these are covered by S-D-Bact-0564-a-S-15/S-D-Bact-0785-b-A-18. Using the GOS dataset, only two sequences from *Firmicutes* are available.

Promising trends could also be observed for the two primer pairs targeting both, *Archaea* and *Bacteria*. In particular, S-D-Arch-0519-a-S-15/S-D-Bact-0785-b-A-18 stands out with high overall coverage (A: 76.5%; B: 83.4%, E: 1.9%), which increases slightly if one mismatch is allowed (A: 81.8%, B: 86.5%, E: 1.9%).

For *Group M*, only 32 sequences of sufficient length were available to re-evaluate the recommended archaeal primer pair S-D-Arch-0519-a-S-15/ S-D-Arch-1041-a-A-18. Thus the *Archaea* primer pairs were excluded from further validation.

With on average 2600 available bacterial sequences for re-evaluating *Group M*, the conditions were slightly better. As in the previous evaluation, several primer pairs show high overall coverage: S-D-Bact-0564-a-S-15/S-*-Univ-1100-a-A-15 proves its suitability with a high domain specific and overall coverage (A: 0.0%, B: 76.2%, E: 0.0%). Overall coverage for *Bacteria* increases up to 80.2%, if one mismatch is tolerated (A: 2.3%, B: 80.2%, E: 0.0%). In contrast, S-D-Bact-0341-b-S-17/S-D-Bact-1061-a-A-17 (A: 0.0%, B: 58.9%, E: 0.0%) fails to match the previous results, which could be a consequence of the specific dataset. Even allowing one mismatch does not achieve satisfying results (A: 0.0%, B: 64.8%, E: 0.0%). At first glance, similar results were obtained for S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21

(A: 0.0%, B: 43.1%, E: 0.0%). However, considering one mismatch the overall coverage significantly increased to A: 58.2%, B: 70.9%, E: 0.0%.

The re-evaluation of the primer pairs based on the GOS dataset (Supplementary Table 27-38) shows that, despite the relatively large dataset size, it still lacks resolution power, especially when considering a specific gene. Unfortunately, the data obtained by other large scale projects, such as the Earth Microbiome Project (163), is of little use for primer evaluation due to their cost effective, but length limited sequencing strategy. Due to the inherent risk of creating chimeric sequences we would not consider assembly a solution to this limitation. Should the error rate of long read sequencing technologies such as PacBio be significantly reduced, data from metagenomic studies relying on these technologies would become a valuable resource for revisiting the primer sensitivity issue. In summary, if a sufficient amount for metagenomic 16S rDNA sequences were available, the previous primer pair recommendations could be confirmed.

Experimental evaluation of the primer pair S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21

The primer pair S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21 (*Group M*) was applied to DNA extracted from a time series of three marine environmental samples at Helgoland Roads. For simplification, we will refer to the obtained reads as '16S pyrotags'. In the course of the MIMAS (Microbial Interactions in MArine Systems) project, metagenomic analysis was performed using marine samples from the same site and time points (140). The results from the metagenomic based diversity studies are used to evaluate the accuracy of each primer pair by comparing the taxonomic classifications.

On average, 59,700 sequences were obtained per sampling occasion, of which 52,400 could be assigned as 16S pyrotags (88.4%) (Supplementary Table 39). The relatively high loss is due to the stringent quality checks used for the identification and taxonomic classification of 16S rDNA fragments. In contrast, metagenome analysis resulted on average in 2,109,000 sequences (140) per sampling occasion, but only 1600 sequences (0.1%) qualified as 16S rDNA gene fragments.

The results of the 16S pyrotag analysis show that the bacterial community is dominated by *Alphaproteobacteria, Bacteroidetes* and *Gammaproteobacteria* (Figure 5A and Supplementary Table 40). According to the *in silico* evaluation, for primer pair S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21 high coverage of these three groups are expected (*Bacteroidetes*: 89.2%; *Alphaproteobacteria*: 81.4%; *Gammaproteobacteria*: 90.6%).

Allowing one mismatch the overall coverage increases to up to 95% for each group. The results from the 16S pyrotags also revealed a succession of the relative abundances. *Bacteroidetes* peaked on 07.04.2009, but were still abundant on 14.04.2009. For *Alphaproteobacteria* more sequences could be detected in winter on the 11.02.09. In contrast, the relative abundance of *Gammaproteobacteria* increased on the 14.04.2009. The same trends were observed in the metagenomes (Figure 5B and Supplementary Table 41) (140). To verify that the results derived from the 16S pyrotags are not an artefact of deep sequencing, the total number of reads was adjusted to smaller subsets of around 2000 sequences by random re-sampling. Detailed analysis of these re-sampled subsets confirmed the results (Supplementary Table 42).

16S pyrotag analysis provides an enhanced resolution up to the group or genus level. Six relatively abundant taxonomic groups and genera (*Formosa, Polaribacter,* SAR11 clade surface 1, NAC11-7 lineage, *Reinekea* and SAR92 clade) have been examined in detail (see Supplementary Fig. 1A and Supplementary Table 43). Noteworthy is the *Formosa* peak on the 07.04.2009 and the presence of *Reinekea* only on 14.04.2009. Both results were supported by diversity studies from the corresponding metagenomes (see Supplementary Fig. 1B and Supplementary Table 44). Again, the re-sampled 16S pyrotag subsets confirmed that the results are not an artefact of deep sequencing (Supplementary Table 45). In addition, it is interesting to note that corresponding metaproteome studies described in Teeling et al. (140) reflect the same succession of the bacterial community on the protein level.

Considering the *in silico* evaluation, S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21 should fail to detect SAR 11 clade surface 1 (0.7%). However, experimental evaluation clearly shows that the primer pair is able to amplify this taxonomic group. This can be explained by the increased coverage of up to 97% if one mismatch is allowed. A closer look at the primer target position of the reverse primer reveals an internal mismatch position towards the 5' end. The results demonstrate that S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21 provides a good representation of the bacterial diversity down to genus and group level and illustrates that an internal mismatch towards the 5' end can be tolerated by standard PCR.

To test the assumption that a suboptimal primer pair might result in a biased picture of the bacterial diversity, S-D-Bact-0008-a-S-16/S-D-Bact-0907-a-A-20 was applied to the same samples. This primer pair was chosen due to its relatively high overall coverage (A: 0.0%, B: 75.1%, E: 0.0%) but distinctly lower phylum spectrum. Based on the *in silico* evaluation it should fail to detect 18 bacterial phyla (*Aquificae,* BD1-5, BHI80-139, *Chlamydiae, Dictyoglomi,* EM19, *Lentisphaerae,* SM2F11, *Thermotogae*, *Tenericutes, Verrucomicrobia,*

WCHB1-60, and Candidate divisions TM7, WS6, OD1, SR1 and OP11). With relatively high coverage of *Bacteroidetes* (77.6%), *Alphaproteobacteria* (71.3%) and *Gammaproteobacteria* (80.5%) *in silico* evaluation and experimental data confirm that this primer pair is able to detect the same dominant taxonomic groups (Supplementary Fig. 2 and Supplementary Table 46). However, in comparison with the 16S pyrotags generated with S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21 and metagenome studies *Alphaproteobacteria* appear to be more abundant throughout all samples. *Bacteroidetes*, on the other hand, are underrepresented. A similar bias can be found on the group level (Supplementary Fig. 3 and Supplementary Table 47). Use of this primer pair indicates a higher relative abundance of *Alphaproteobacteria* SAR11 clade surface 1 as well as NAC11-7 lineage on 07.04.2009 and 14.04.2009. In turn, particularly the genus *Formosa* is less prominent. This is in line with the results from the *in silico* evaluation, which shows that S-D-Bact-0008-a-S-16/S-D-Bact-0907-a-A-20 only detects 52.9% of the *Formosa* sequences. Even one allowed mismatch results only in an increase of 9% up to 61.9%. A closer look reveals a mismatch of the reverse primer towards the 3'end for several *Formosa* sequences.

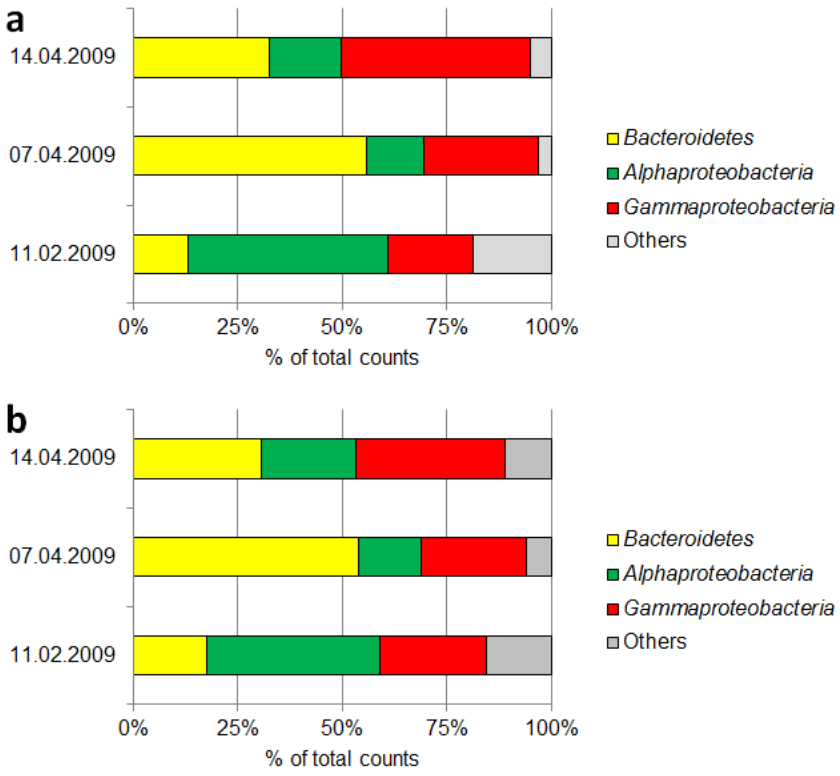


Figure 5: Taxonomic distribution of 16S rRNA gene sequences gained from a time series of three different surface water samples at Helgoland Roads in the North Sea, (a) 16S pyrotags generated from PCR and sequenced with Roche's 454 pyrosequencing (relative abundance, % of total counts) (b) 16S sequences gained from metagenome studies (relative abundance, % of total counts).

Although S-D-Bact-0008-a-S-16/S-D-Bact-0907-a-A-20 is able to detect all major groups, a bias in the relative abundances as well as community structure is clearly confirmed by the experimental data (Fig. 1 and Supplementary Fig. 1-3). This supports our assumption that the overall coverage need always to be considered in combination with the phylum spectrum. Detailed analysis of the mismatch position should also be taken into account. Nevertheless, the experimental results strongly indicate that *in silico* evaluation can serve as a guideline for choosing the most suitable primer pair.

2.5. Conclusion

The advent of new sequencing methods has been a paradigm shift for molecular ecology and especially microbial diversity analysis using marker genes. The rapid adoption of the new techniques caused a backlog in proper evaluation of the primers used for diversity surveys. Our study shows that even commonly used single primers exhibit significant differences in overall coverage and phylum spectrum. Consequently, primer pairs need to be carefully selected to avoid accumulative bias. Out of the 175 primers and 512 primer pairs checked, only 10 can be recommended as broad range primers. Although none of them are perfect, and especially for the *Archaea* we recommend the design of additional primers, the experimental validation shows that a good combination of primers approximate PCR free metagenomic approaches with respect to community structure and relative abundances. The results confirm that single internal mismatches, when located towards the 5' end, are tolerated in the amplification process. Re-inspection of the primers using GOS metagenomes was found to be a reasonable approach for determining possible primer bias in the public rDNA repositories. However, the incomplete phylum spectrum as well as the comparatively small dataset size with respect to 16S rRNA genes in the GOS metagenomes did not allow for an in-depth re-evaluation. For example, *Group M* primer pair S-D-Bact-0341-b-S-17/S-D-Bact-0785-a-A-21, which we recommended based on the SSURef 108 NR results, fails to detect major groups in the GOS dataset, yet excels in the experimental evaluation. This demonstrates the validity of using comprehensive, non-redundant datasets like the SILVA SSURef 108 NR for in depth evaluation of probes and primers. We would like to note that the SILVA project has prepared an online service for this purpose at www.arb-silva.de/search/testprime, which is modelled after our evaluation method and allows inspection of per-taxon coverages for individual primer pairs. Furthermore, all primers, including bibliographic information and information on specificity and overall coverage, have been added to probeBase. The availability of the evaluated primers in a central and publically accessible repository plus the online primer

evaluation tool should facilitate the search for, and the evaluation and selection of, suitable primers in future studies.

Supplementary Data

Supplementary Data are available at NAR Online and http://www.arb-silva.de/download/archive/primer_evaluation/: Supplementary Fig. 1-3 and Supplementary Table 1-52.

Acknowledgments

We acknowledge Jack A. Gilbert (Argonne National Laboratory, Argonne, IL, USA), Bernhard M. Fuchs (Max Planck Institute, Bremen, Germany) and Christine Klockow for critical discussion of this manuscript. E. Karamehmedovic and M. Meiners for helping with the laboratory work. G. Gerds and A. Wichels from the Alfred Wegner Institute (Bremerhaven, Germany) for supporting and performing the water sampling. Hannah Marchant, Elizabeth Robertson, Mira Okshevsky and Mario Schimak for critical reading of the manuscript.

3. Chapter

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

Authors: Hanno Teeling, Bernhard M. Fuchs, Dörte Becher, Christine Klockow, Antje Gardebrecht, Christin M. Bennke, Mariette Kassabgy, Sixing Huang, Alexander J. Mann, Jost Waldmann, Marc Weber, Anna Klindworth, Andreas Otto, Jana Lange, Jörg Bernhardt, Christine Reinsch, Michael Hecker, Jörg Peplies, Frank D. Bockelmann, Ulrich Callies, Gunnar Gerdts, Antje Wichels, Karen H. Wiltshire, Frank Oliver Glöckner, Thomas Schweder, Rudolf Amann

Status: published in Science. 2012; 336(6081):608-611

Contribution: involved in the sampling procedure, based on the previous evaluation design and performance of 16S pyrotag analysis

3.1. Abstract

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.

3.2. Manuscript

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(164), yet they are responsible for ~19% of the oceanic NPP (165) and globally account for 80% of organic matter burial and 90% of sedimentary mineralization (164). Heterotrophic members of the picoplankton—mostly *Bacteria*—reprocess about half of the oceanic NPP in the so-called 'microbial loop' (166). The bulk of this bacterioplankton biomass is free-living, but up to 20% is attached to algae or particles (167).

The bacterial response to coastal phytoplankton blooms has been almost exclusively studied in microcosm/mesocosm experiments (84,168-170) or with limited resolution in time and biodiversity in situ (171-173). 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 (Figure 6A) were dominated by *Alphaproteobacteria* (41 to 67%), composed roughly of two-thirds SAR11 clade and one-third *Roseobacter* clade (Figure 6B and Fig. S4B). SAR11 consisted almost exclusively of subgroup Ia (*Candidatus Pelagibacter ubique*) (table S4). This composition changed as the spring phytoplankton bloom commenced (further information is available as supplementary materials on *Science* online). 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 (Figure 6C 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 (Figure 6D 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).

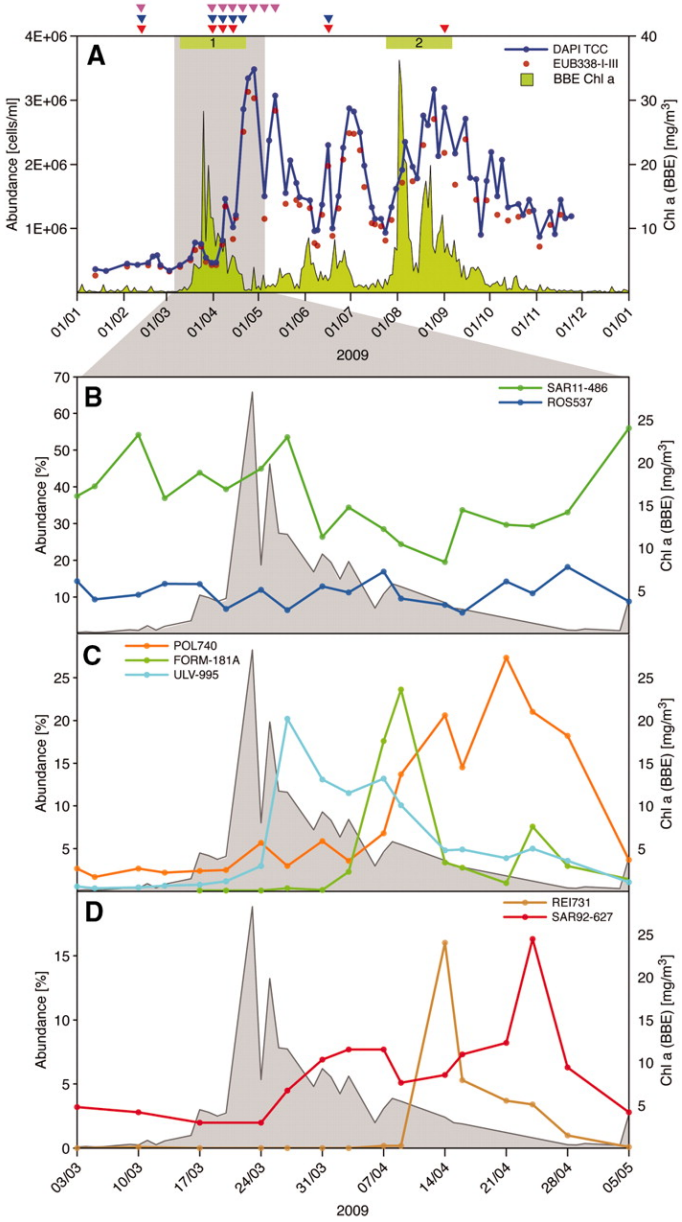


Figure 6: Abundances of major bacterial populations during the bacterioplankton bloom as assessed by CARD-FISH. (A) Chlorophyll a (Chl a) concentration (measured with a BBE Moldaenke algal group analyzer), 4′,6-diamidino-2-phenylindole (DAPI)–based total cell counts (TCC), and bacterial counts (probe 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 16S 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).

Metagenomes were partitioned into taxonomically coherent bins (taxobins, fig. S5A) and then used for identification, annotation, and semiquantitative analyses of the metaproteome data (further information is available as supplementary materials on *Science* online). This allowed the investigation of shifts in gene content and expression within dominating bacterial populations (table S7).

A pronounced peak in the abundance of carbohydrate-active enzymes [CAZymes (174)] accompanied the bacterial succession (fig. S5B). CAZyme frequencies and expressions were taxonomically distinct (Figure 7 and Figure 8). 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 (175,176). *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 (Figure 8) and was dominated by *Flavobacteria*. Expressed sulfatases were found in the *Polaribacter* taxobin, which corroborates recent reports of high numbers of sulfatases in *Polaribacter* (177). In contrast, glycoside hydrolases for decomposing nonsulfated laminarin (GH16, GH55, and GH117) had their expression maxima earlier during the initial algal die-off phase.

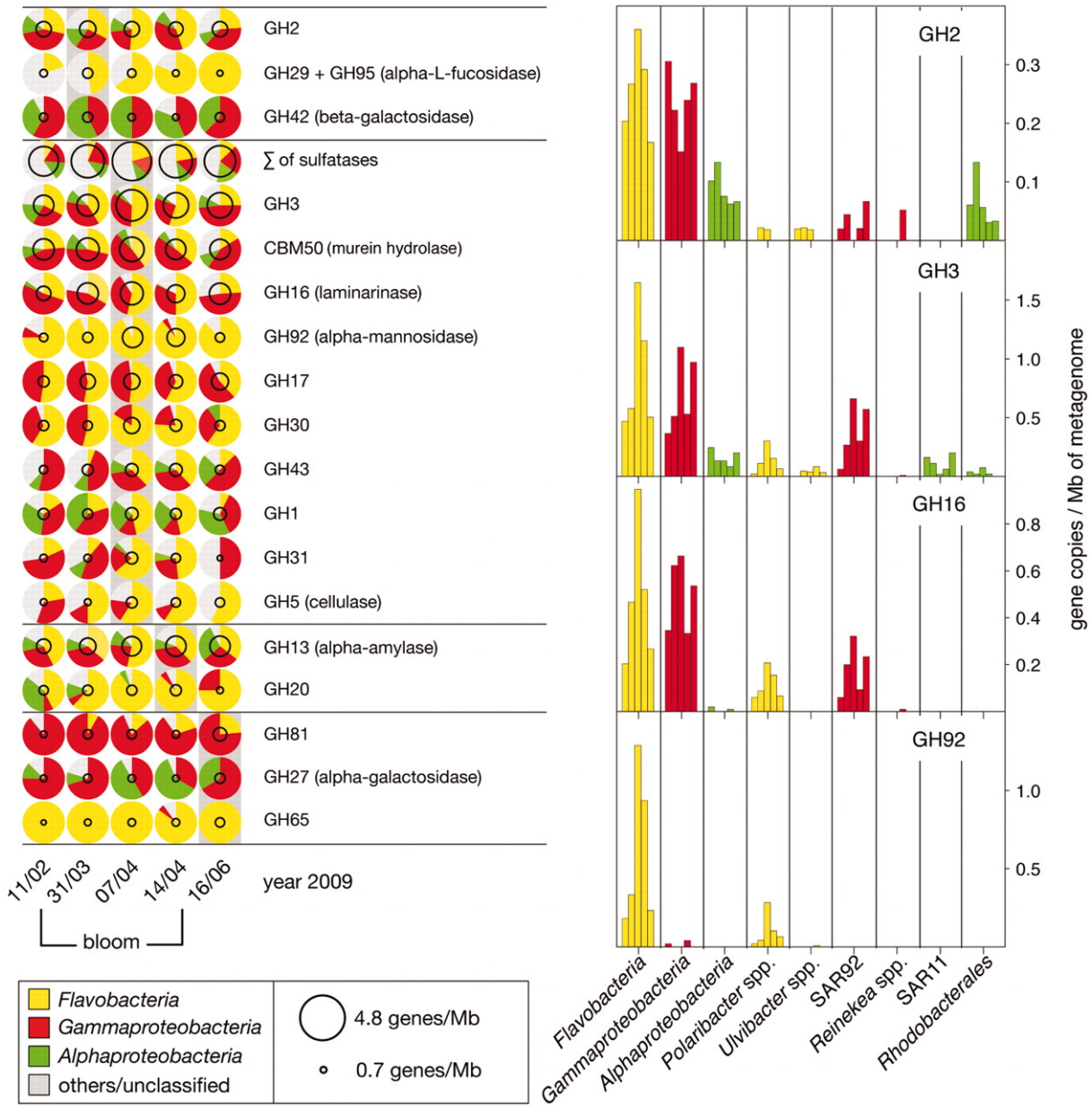


Figure 7: Abundances of CAZymes with relevance for external carbohydrate degradation. (Left) Copies of 20 CAZymes per megabase of metagenome sequence with class-level taxonomic classifications (further information is available as supplementary materials on *Science* online). Maximum abundances are highlighted 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).

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 (Figure 9A).

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 (Figure 9A). TBDTs, originally thought to be restricted to complexed iron(III) (178) and vitamin B12 uptake, allow uptake of compounds that exceed the typical 600- to 800-dalton substrate range of normal porins (179,180). Within Bacteroidetes, TBDTs are often colocalized with carbohydrate degradation modules (fig. S6) (177,181-183), and thus the substrate spectrum of these transporters may be much wider than anticipated (184), 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 (185). 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.

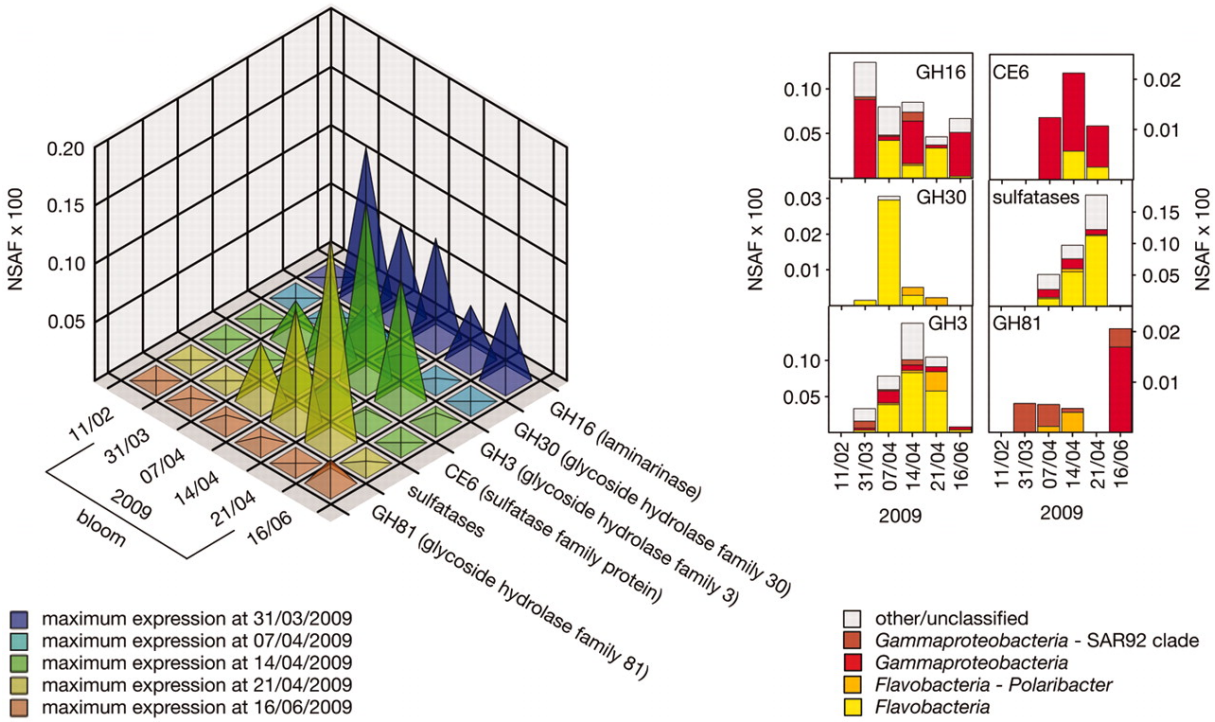


Figure 8: Expression of CAZymes with relevance for external carbohydrate degradation; the proteome data were analyzed in a semiquantitative manner based on normalized spectral abundance factors (NSAFs) (further information is available as supplementary materials on *Science* online).

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 (Figure 9A).

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 (186-188). Our data confirmed constitutive proteorhodopsin expression and transporter components as the most abundant expressed proteins in the SAR11 clade, which corroborates previous findings (189). Members of the metabolically diverse, opportunistic alphaproteobacterial *Roseobacter* clade (190-192) exhibited LMW transporter expression levels that exceeded those of SAR11 (Figure 9A). 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 (193); 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 systems in various bacterial taxobins increased over the progression of the bloom (Figure 9B). *Gammaproteobacteria* and SAR11 tended to use ABC-type phosphate transporters, as discovered in earlier studies (189), 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 (170). *Bacteroidetes* have also been identified as major bacteria attached to marine snow(131,194), which agrees with their presumed role as fast-growing r strategists with specialization on the initial attack of highly complex organic matter(177,181,195). Hence, algal blooms lead to a multifold increase of colonization surfaces for *Bacteroidetes*, which respond with increased production of exoenzymes (196). After algal lysis, *Bacteroidetes* are the first to profit.

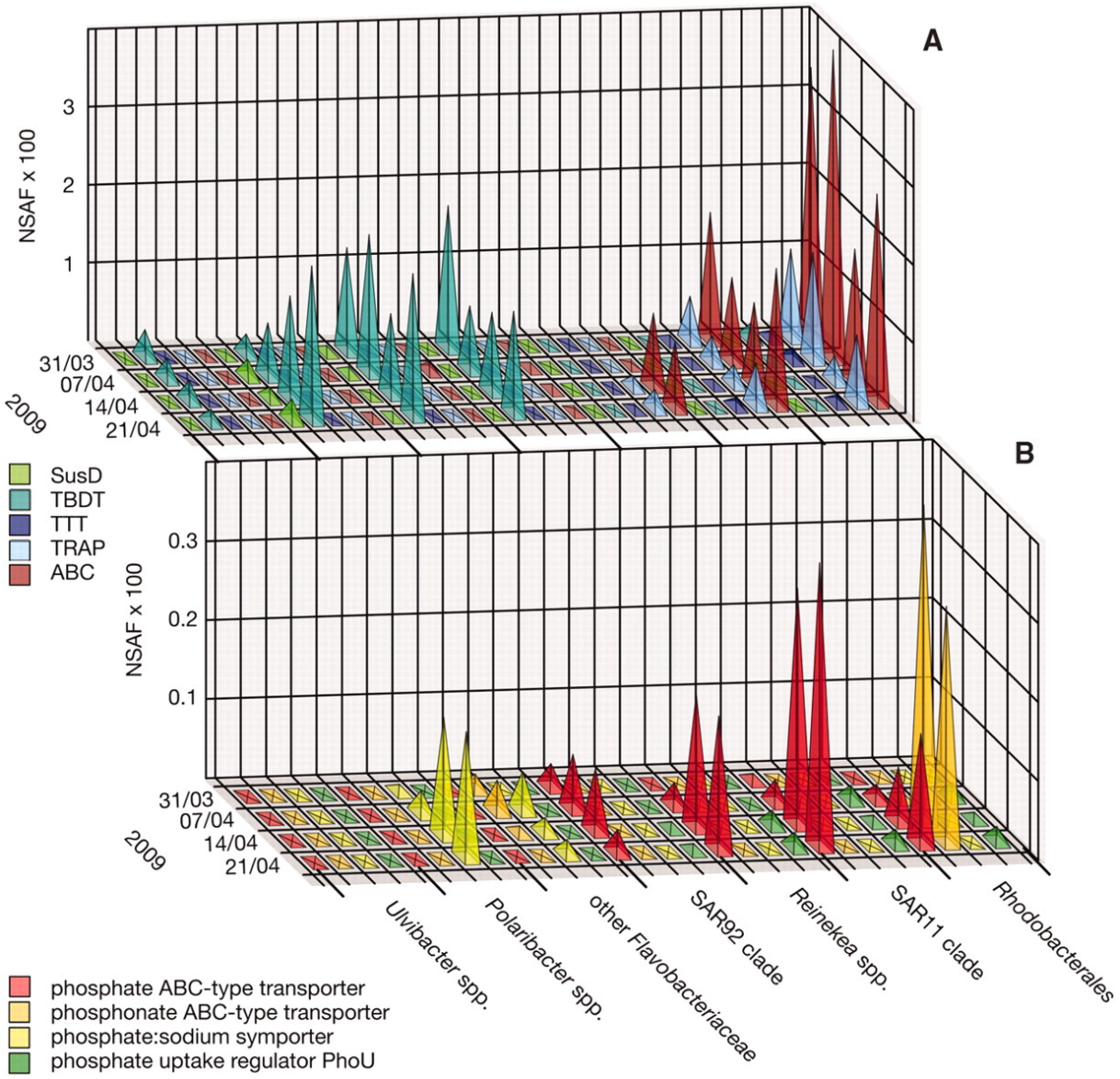


Figure 9: Transporter components and phosphorus acquisition proteins of dominant taxa during the bacterioplankton bloom. (A) Expression of transporter components: starch utilization SusD-family proteins (SusD), TBDTs, TTTs, TRAPs, and ABCs. (B) Expression of proteins involved in phosphorus acquisition.

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 (197-199) that found broad generalist substrate spectra. The increase of alphaproteobacterial *Roseobacter* clade RCA during this phase might also be attributed to the *Roseobacter's* opportunistic life-style (190) and is consistent with previous findings of free-living RCA phylotypes in the German Bight during diatom blooms (200).

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. 6, C and D). At this time, *Polaribacter* and *Formosa* dominated the particle/algae-attached fraction (table S8). Hence this phase with high sulfatase expression (Figure 8) reflected another change of ecological niches (further information is available as supplementary materials on *Science* online).

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 (201) and thus can tackle the 'paradox of the plankton' (202) , which is how these species evade extinction by direct competition in a seemingly homogeneous habitat with limited resources.

Supplementary Data

Supplementary Data are available at Science online:
http://www.sciencemag.org/content/336/6081/608/suppl/DC1

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).

4. Chapter

Comparative metatranscriptome analysis of a diatome-induced North Sea bacterioplankton bloom

Autors: Anna Klindworth, Alexander Mann, Sixing Huang, Jost Waldmann, Jörp Peplies, Christian Quast, Christine Klockow, Hanno Teeling, Frank Oliver Glöckner

Status: draft to be submitted to Microbial Genomics

Contribution: design and performance of laboratory experiments, analysis of data and writing the manuscript

4.1. Abstract

Metatranscriptomics and metaproteomics are two different approaches for studying active genes of microbial cells and their responses to environmental changes *in situ*. Using metatranscriptomics, we investigated the response of subsurface bacterioplankton communities from Helgoland Roads in the North Sea to a diatom-dominated spring phytoplankton bloom. Afterwards, we compared these data to metaproteome data of the same samples that we published previously. Metatranscriptomics could enhance the resolution of the metaproteome results by providing additional taxonomic and functional information down to the genus level. Simultaneous detection of cDNA derived from mRNA and rRNA revealed that in particular *Roseobacter* and *Reinekea* clade members adapted rapidly to changes in substrate composition the course of the bloom. High rRNA expression levels and fast mRNA degradation characterized these bacteria, which probably allows them to quickly react successfully compete for specific algae-derived substrates in the moment they become

available. Taxonomically distinct mRNA expression of membrane transporters and carbohydrate active enzymes (CAZymes) further supported such distinct nutrient utilisation strategies within different clades of *Flavobacteria*, *Alphaproteobacteria* and *Gammaproteobacteria*. This fortifies the hypothesis that during the investigated spring diatom bloom changes in algal substrate composition provided distinct ecological niches in which specifically adapted bacterial clades to grow.

Introduction

Marine microorganisms represent a valuable resource for new promising gene functions, enzymes and bioactive substances (203,204). In this respect it is of fundamental interest to extend our knowledge about the genes and functions of marine microbes. Unfortunately, investigations on the molecular level are hampered by our inability to grow the majority of marine microorganisms in pure cultures or under laboratory conditions (27). With the advent of high throughput sequencing technologies meta-'omics' approaches have enabled culture independent *in situ* studies of marine microorganisms without prior cultivation (205-207). In particular metagenomics has become the standard tool for the analysis of marine bacterial communities (140,208). However, it can neither reveal whether the sequenced DNA comes from vital cells, nor whether the obtained genes are expressed under the actual environmental conditions. Such questions can be addressed by cultivation independent gene expression analysis of bacterial communities, such as metatranscriptomics and metaproteomics. These approaches shed different lights on gene expression and must be regarded as complementary. Both approaches have substantially advanced environmental genomics, which is reflected in numerous recent studies of a variety of marine habitats that have addressed either the metatranscriptome (75-77,81-84) or metaproteome (140,189,209,210).

Metatranscriptomics of total RNA allows combined taxonomic and functional investigations of the sample with a single technique and the discovery and functional analyses of putative small regulatory RNAs (sRNA) (89). Cost-effective high throughput sequencing technologies allow large transcriptome (cDNA) data sets, which constitutes another major advantage of the metatranscriptome approach. On the down side, metatranscriptomics does not reflect all regulatory processes in the bacterial cell such as post-transcriptional, translational and post-translational regulation (115,116). Likewise, metatranscriptomics remains technologically challenging due to the short life span of prokaryotic mRNA (95-97) and the low mRNA to rRNA transcript ratio within a total RNA sample (99). Unlike mRNA, proteins are much more stable (211), allowing metaproteomics to provide a better and more accurate determination of

the abundantly expressed functional genes. However, metaproteomics is hampered by the required high amount of proteins, the broad range of protein expression levels, incomplete gel electrophoretic protein separation and high-cost of mass spectrometry protein identification (212). Moreover, protein identification requires a corresponding metagenome, and in comparison with metatranscriptomics results in a much smaller dataset (117). Hence, no single approach alone can fully unravel the complexities of the functional dynamics of microbial communities (116). Instead, an integrative analysis of both currently constitutes the best approach for studying gene expression in a microbial community (115). First, a combined approach increases the confidence at which dominantly expressed genes are detected. Second, high throughput metatranscriptome datasets provide comprehensive information with high resolution power and greater taxonomic accuracy.

Recently, we applied an integrated multi 'omic' approach to investigate the bacterioplankton response to a diatom-dominated phytoplankton bloom in the North Sea (140). The combination of metagenomics and metaproteomics studies uncovered distinct expression patterns in specific clades of *Flavobacteria*, *Gammaproteobacteria*, and *Alphaproteobacteria*. The results revealed a taxonomic specialized successive decomposition of algal-derived organic matter that could be linked to different environmental lifestyles. In the present study, we first compare the 16S ribosomal gene (16S rDNA) encoding reads derived from directly sequenced total RNA (cDNA), pyrotags and metagenomes for three selected timepoints at the sampling site Helgoland Roads in the German Bight of the North Sea. This provided not only information on the presence and absence of taxa, but for the first time uncovered the metabolically most active members of the bacterioplankton community. Second we tested whether metagenomic, metatranscriptomic, and metaproteomic profiles followed similar patterns.

4.2. Materials and Methods

Sampling site and collection of water samples

Sample collection was carried in the framework of the MIMAS (Microbial Interaction in MArine Systems) project (www.mimas-project.de). Subsurface water was collected on 11th February 2009 and weekly from the 31th of March 2009 until October 2009. Water samples (total volume 360 L) from the Kabeltonne site at Helgoland Roads in the North Sea (54°11.18′N, 7°54.00′E) were collected at a depth of 0.5 m and processed immediately at the Biological Station Helgoland. The water was pre-filtered through a 10 µm and a 3 µm pore-

size polycarbonate filter (142 mm TSTP, Millipore). For harvesting a 0.2-µm-pore-size polyethersulfone filters (142 mm GPWP, Millipore) was used. At each time point 10 L and 15 L of seawater were filtered onto eight filters for RNA and genomic DNA extraction, respectively. All filters were stored at -80°C until future usage. Further details can be found in Teeling et al. (140). In this study, metatranscriptomics and 16S rDNA analysis from total RNA as well as community DNA were performed using samples from: 11.02.2009, 31.03.2009 and 14.04.2009.

DNA Extraction

Genomic DNA was directly extracted from filters as described in Zhou et al. (148) with the following modifications: all extraction steps were performed with 50 µl proteinase K (10 mg/ml), and after isopropanol precipitation, pelleted nucleic acids were obtained by centrifugation at 50,000 g for 30 min at room temperature. The genomic DNA was stored at -20°C until PCR amplification and metagenomic sequencing were carried out.

Amplification of 16S rDNAs

PCR reaction for 16S rDNA gene amplification the was carried as in two previous studies (140,213). The forward primer was S-D-Bact-0341-b-S-17, 5'-CCTACGGGNGGCWGCAG-3' (149), and the reverse primer S-D-Bact-0785-a-A-21, 5'-GACTACHVGGGTATCTAATCC-3 (149). A second primer pair (S-D-Bact-0008-a-S-16/S-D-Bact-0907-a-A-20) was evaluated, but not used in this study. For details please refer to Klindworth et al. (213).

RNA extraction and mRNA enrichment

Filters were incubated in 10 mL of Solution D (214). The suspension was incubated for 5 min at room temperature. Cells were lysed by bead-beating (lysing matrix B, material: 0.1 mm silica spheres; MPBiomedicals, Berlin, Germany) applying a FastPrep 24 automated homogenizer (MPBiomedicals). Three steps of 30 s (speed: 6 m/s) were performed, while cooling the tubes on ice between beadbeating steps. After the third step, the beadbeater tubes were incubated on ice for an additional 10 min. Afterwards, the tubes were centrifugated at 4 °C for 10 min (5415 C, Eppendorf, Hamburg, Germany; 13,200 rpm, rotor: FA-45-24-11). Supernatants (1000 µl each) were transferred into RNase-free, sterile 1.5 mL Eppendorf cups. 200 µL of ice-cold chloroform was added per sample. Suspensions were thoroughly mixed by

vortexing for 20 s, followed by a 2 min incubation step at room temperature (RT). A further centrifugation step was carried out (4 °C, 15 min, 13200 rpm). The aqueous, upper phase was transferred into new, RNase-free and sterile Eppendorf cups. 1 mL of 100% isopropanol was added, followed by incubation at –20 °C for one hour. After the incubation, a 30 min centrifugation was performed (4 °C, 13200 rpm). The supernatants were discarded and pellets were washed twice in 75% (v/v) ethanol. Dried pellets were dissolved in 50-100 µl RNase-free water. Extracted RNA was cleaned using the RNeasy MinElute clean-up kit (Qiagen, Hilden, Germany) following the manufacturer's instructions with the modification of 700 µl instead of 250 µl 96% (v/v) Ethanol in the second step. The eluted RNA was treated with TURBO™ DNase (Ambion, Austin, TX, USA) following the manufacturer's instructions to remove DNA contaminations. The concentration and quality of eluted RNA was determined using a NanoDrop® spectrophotometer (ThermoFisher Scientific Inc., Wilmington, MA, USA). The amount and quality of extracted and cleaned up RNA was also documented by RNA agarose gelelectrophoresis. Samples for 16S rDNA analysis from total RNA were immediately used for cDNA synthesis. Metatranscriptomic samples (31.03.2009 and 14.04.2009) had to undergo mRNA enrichment prior to cDNA synthesis using the mRNA only Prokaryotic mRNA isolation kit (Biozym Scientific, GmbH, Hessisch Oldendorf, Germany) and MICROB/Express/™ Bacterial mRNA Enrichment Kit (Ambion, Austin, USA). The 11.02.2009 winter sample was excluded from mRNA enrichment and was subject to immediate cDNA synthesis due to the low biomass. Instead this sample was used in a combined approach of functional (metatranscriptomics) and taxonomic (16S rRNA) analysis from cDNA.

Synthesis of cDNA

Synthesis of cDNA was carried out using SuperScript® Direct cDNA Labelling System (Life Technologies, Darmstadt, Germany). The first strand cDNA synthesis reaction was followed by a second strand cDNA synthesis with Polymerase (30 U), 10 x strand buffer and RNase H (1 U) (Fermentas, St. Leon-Rot, Germany). The reaction was carried out in total volume of 100 µl at 15 °C for two hours. Blunt ends were generated with T4 DNA polymerase (12.5 U) (Fermentas) at 15°C for 5 min. The reaction was terminated with 0.5M. EDTA. The cDNA was purified with the QIAEX II Gel Extraction Kit (Qiagen). The quantity and quality of the extracted cDNA were analyzed using ND-1000 spectrophotometer (NanoDrop Technologies, Wilmington, MA, USA) and by agarose gel electrophoresis. The cDNA was stored at -20 °C until future use.

Sequencing

The pyrosequencing was carried out on a 454 FLX Ti pyrosequencer (Roche/454 Life Sciences, Branford, CT, USA) at LGC Genomics (LGC Genomics GmbH, Berlin, Germany). Library preparation and sequencing were performed according to the manufacturers' protocols. In brief, cDNA from total RNA and mRNA-enriched samples were checked for quality on a 2% agarose gel. Afterwards, 500 ng of each sample were used to construct the sequencing library. No size selection of fragments was conducted in order to retain potential small RNAs. The fragments were subjected to end repair and polishing. An extra A was added to the ends of the fragments, and Roche Rapid Library adaptors were ligated to the fragments as described in the Roche Rapid Library Preparation Manual for the GS FLX Titanium Series, October 2009, Rev. Jan. 2010 (Roche/454 Life Sciences). After subsequent emulsion PCR the fragment libraries were processed and sequenced according to the Roche protocols. The resulting sequences were processed using the standard Roche software for base calling, trimming of adaptors and quality trimming (Genome Sequencer FLX System Software Manual version 2.3; Roche/454 Life Sciences). For cDNA synthesized from untreated total RNA samples 1/8 picotiter plate (PTP) was sequenced for each sample. For the two enriched mRNA metatranscriptomic datasets a complete PTP was sequenced for each sample. Sequences were submitted to the INSDC (EMBL-EBI/ENA, Genbank, DDBJ) with accession numbers xxx, yyy, zzz. Genomic DNA from metagenome studies as well as PCR amplified DNA fragments were sequenced within two previous studies (140,213). For PCR amplified DNA fragments, per sample two distinct PCR reactions were sequenced on 1/8 PTP (213). These Sequences are available from INSDC with accession number ERP001031. For metagenomics, 2.5 PTPs (11.02.2009), 2 PTPs (31.03.2009) and 4 PTPs (14.04.2009) were sequenced per sample (140). Sequences are available from INSDC with accession number ERP001227.

Identification and taxonomic classification of 16S rDNA fragments

Unassembled 16S rDNA reads from total RNA, amplified pyrotags and metagenomes were processed by the SILVA bioinformatic pipeline (36) using SINA (143). Details are described in Klindworth et al. (213).

Processing of metatranscriptome data

The metatranscriptome reads identified as rDNA by the SILVA pipeline (36) were excluded. The remaining metatranscriptome reads were mapped with SSAHA2 (215) onto the MIMAS annotated and taxonomically classified metagenome data of the MIMAS project (140), in order to assign read taxonomy and protein function. The best KEGG (216), Pfam (217) and CAZY (174) hits, with e-values above 10 e^{-6} were used for comparison of the metatranscriptome data. Because of multiple mappings of the metatranscriptomic reads to individual genes, a consensus of the results has been carried out on the taxonomic and functional level.

Metatranscriptomes were normalized for comparision. Due to high numbers of hypothetical proteins, metatranscriptomes were normalized based on the genes with known Pfam domain functions (11.02.2009: 39.012 hits; 31.03.2009: 39.518 hits; 14.04.2009: 33.215 hits).

Data overview

This comprehensive analysis uses data from different experimental approaches and studies. An overview of the respective datasets is provided in Table 2. 16S rDNA reads from three different experimental approaches were used for taxonomic profiling: Sequences derived from untreated cDNA, PCR amplified 16S pyrotags and metagenomic 16S rDNA fragments, respectively. For functional analysis, cDNA reads from metatranscriptomes were compared with the outcome from previous metagenome and metaproteome studies (140). All of these data is available for all three sampling dates (11.02.2009, 31.03.2009 and 14.04.2009), which amounts to a total of 18 different datasets. For simplification, each dataset has been assigned with a 'working title'.

Table 2: Overview of the datasets used in this study and their corresponding reference

Experimental approach	Type of sequence	'working title' used in this study	Reference
taxonomic profiling	16S rDNA from directly sequenced cDNA*	16S cDNA	this study
	16S rDNA from PCR amplified pyrotags	16S pyrotags	(213)
	16S rDNA from metagenomes	16S metagenome	(140)
functional analysis	protein coding sequences from cDNA	metatranscriptome	this study
	protein coding sequences from genomic DNA	metagenome	(140)
	expressed protein sequences	metaproteome	(140)

*cDNA was synthesized from an untreated total RNA sample

4.3. Results and Discussion

Taxonomic profiles of the microbial communities

In this study, the 16S rDNA fraction of the of the total RNA was compared to the previously described bacterial community structures at the island of Helgoland in the German Bight (140). For this purpose 16S rDNA sequences of selected sampling points (11.02.2009, 31.03.2009 and 14.04.2009) were retrieved from directly sequenced cDNA, and compared to 16S rDNA from metagenomes (140) and PCR amplified pyrotags (140,213). For simplification we will refer to the working titles '16S cDNA', '16S metagenome' and '16S pyrotags', respectively, as described in Table 2.

Although 16S cDNA and 16S pyrotags exhibit on average 25 times larger datasets than 16S metagenomes, comparison of the dominant community members is feasible. Our previous study showed that the results derived from the larger dataset are not an artefact of deep sequencing, and did not infringe on the comparability of the resulting taxonomic resolution (213).

The results gained from 16S cDNA revealed that *Alphaproteobacteria*, *Gammaproteobacteria* and *Flavobacteria* appear to be the most active members of the bacterial communities (Figure 1a). 16S metagenome and 16S pyrotags showed the same trend as demonstrated in Figure 1b-c. In the 16S cDNA winter sample (11.02.2009) the bacterial community was slightly dominated by *Alphaproteobacteria* (Figure 10a), composed mainly of the SAR11 clade and some *Roseobacter* clade members (Supplementary Figure 1a). Interestingly, in the 16S pyrotags and 16S metagenome far more reads could be assigned to *Alphaproteobacteria* and in particular to SAR11 clade members. 16S cDNA sequencing identified many SAR11 to consist of *'Candidatus* Pelagibacter', whose well-studied representative *Pelagibacter ubique* HTCC 1062 encodes one of the smallest known genomes with no duplicate gene copies and just one 16S rRNA gene (218). Taking into account that metabolically active bacteria contain more expressed ribosomal RNA than latent or starved cells (219), we believe that the low amount of 16S cDNA reads indicates low activities as a response to low nutrient conditions in winter. With an assumed rRNA operon copy number of one the rather high amount of 16S reads in the pyrotags and metagenomes might only reflect the high occurrence of SAR11, rather than high activity.

With the occurrence of the spring algae bloom (31.03.20012 and 14.04.2009), all three methods indicate a change in the community structure resulting for example in an increase of

Flavobacteria with high abundances of *Formosa* and *Polaribacter* species (Supplementary Figure 1a-c). The boost of 16S cDNA rates of *Formosa* and *Polaribacter* in early spring suggest higher metabolic activity and supports the assumption that carbohydrate degrading *Flavobacteria* are the first to benefit from the algal substrate availability (140).

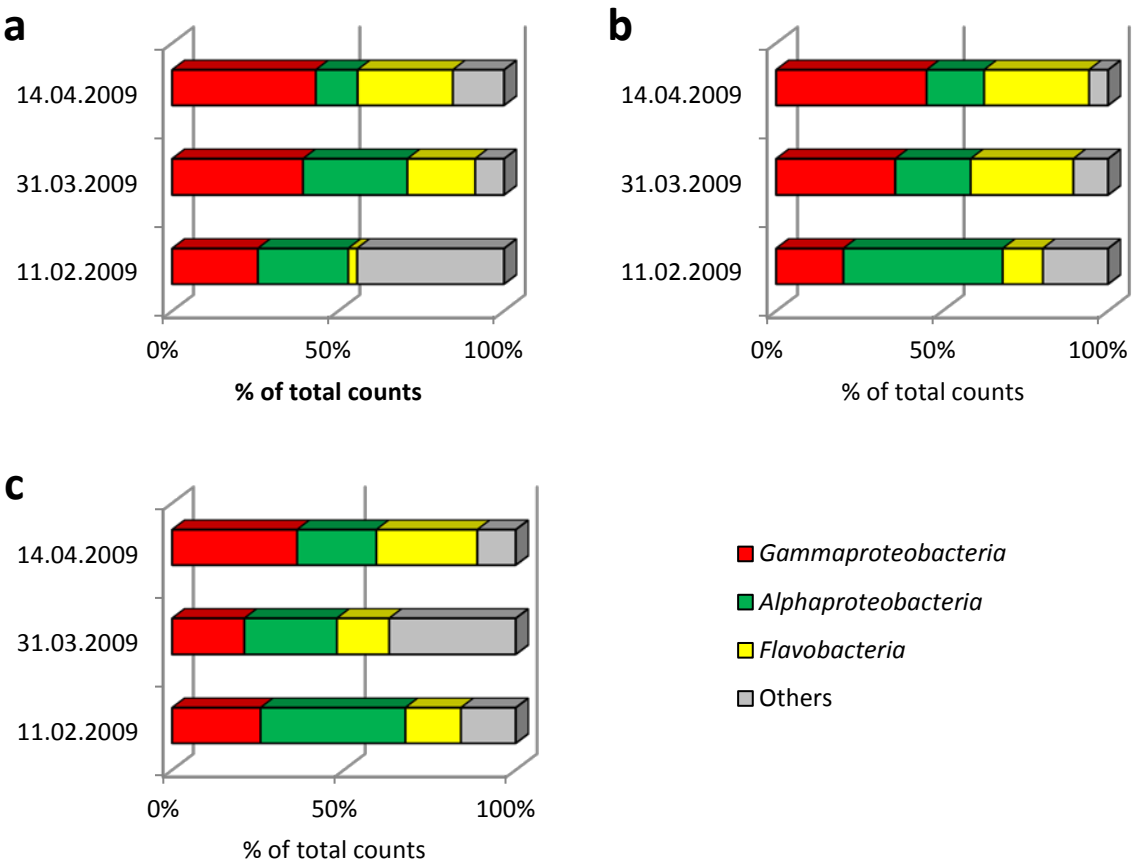


Figure 10: Taxonomic profile of three dominant taxonomic groups. 16S rDNA reads were gained from a) directly sequenced cDNA (16S RNA), b) PCR amplified pyrotags (16S pyrotags) and from c) metagenome (16S metagenome).

In addition, detailed cDNA analysis revealed that the *Roseobacter* clade appeared to be metabolically very active in early spring (Supplementary Figure 1a). Members of this group contain one to five rRNA operons per cell (190), which most likely allows them to respond rapidly to resource availability (220). The distinct 16S cDNA peak is in line with the hypothesis of Giebel at al., who suggested, that *Roseobacter* adapt readily to phytoplankton bloom dynamics (200).

Gammaproteobacteria showed a constant increase in activity and abundance (Figure 1a-c). SAR92 clade and *Reinekea* (Supplementary Figure 1a-c) were identified as two dominant members. SAR92 clade appeared to be very active with a distinct peak of 16S cDNA on the

31.03.2009 sample suggesting that they benefit from the early spring bloom. *Reinekea* appeared 'out of the blue' to gain high abundances within only a week (for details please refer to Teeling et al. (140)). The sudden appearance and activity of *Reinekea* sequences on the 14.04.2009 could be confirmed by the 16S cDNA dataset. Members of this group have up to four rRNA copy numbers, which suggests that they are also prepared to react rapidly to changing environmental conditions.

Functional profile of the bacterial community

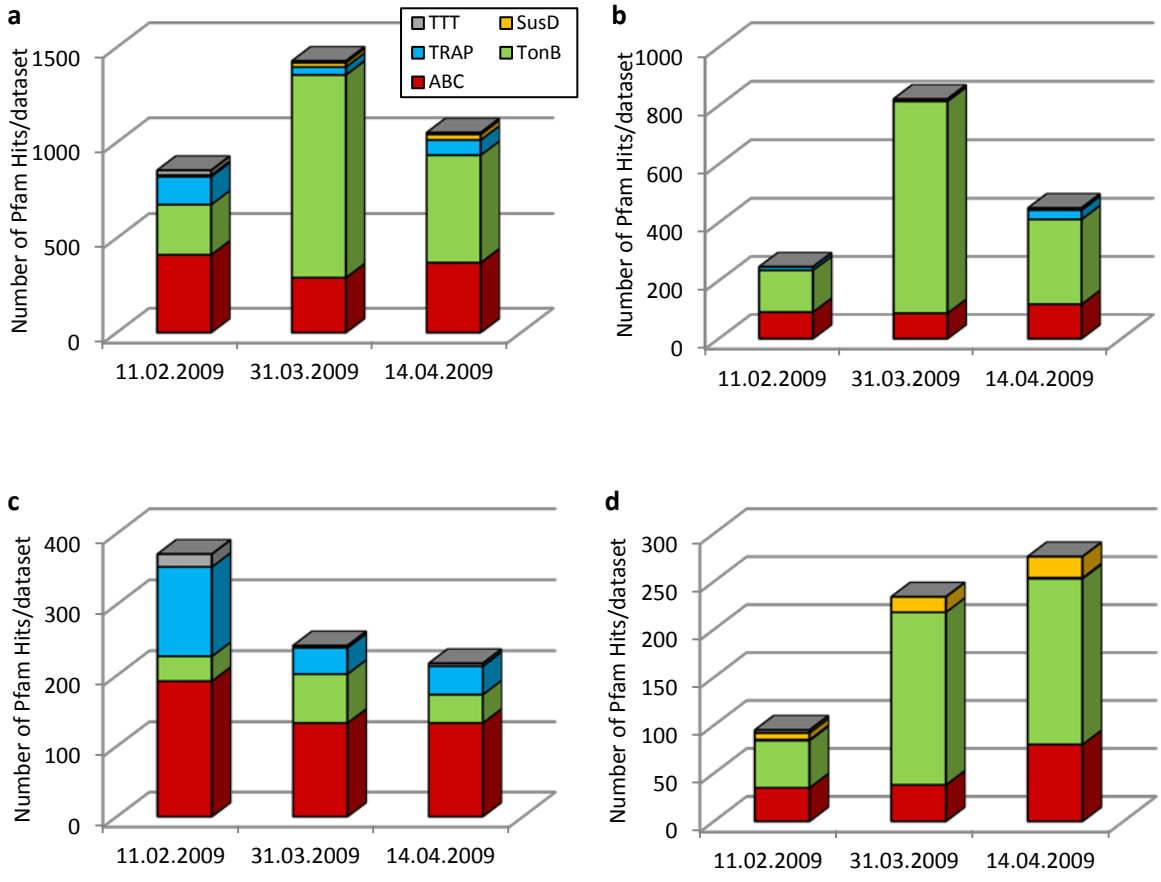
Three metatranscriptome datasets with an average amount of 1,139,553 reads per sampling day were obtained. 80% (11.02.2009) – 91% (31.03.2009) of the reads could be mapped to the metagenome data. 58% (11.02.2009) – 69% (31.03.2009) of all sequences could be assigned onto ORFs within the metagenome datasets. Taxonomic assignment of the reads reflects the taxonomic profile described in the previous section. The core of the metabolic active member includes *Gammaproteobacteria*, *Alphaproteobacteria* and *Flavobacteria* (Supplementary table 2). Not surprisingly the most abundant transcripts with known function could be assigned to housekeeping genes such as, but not limited to, elongation factors, DNA gyrase and sigma factors, indicating fit and active microbial cells. Moreover, stress-induced chaperonin proteins were also found among the most abundant reads. This particular stress-response could be a result of the intense filtering time as previous suggested by Gilbert et al. (76).

Pfam annotations yielded significant numbers of membrane transporters (Figure 11a) reflecting differences in nutritional ecological strategies of the dominant bacterial groups. Among those abundant transcripts were genes encoding for TonB-dependent transport systems (TBDT), starch utilization system proteins (SusD), and other low-molecular weight (LMW) transporters such as ATP binding cassette (ABC), tripartite ATP independent (TRAP) and tripartite tricarboxylate transporters (TTT).

Initially substrates of TBDT were thought to be restricted to iron and vitamin B12 (184,221). However, recent analysis of the genetic content near TBDTs in marine bacteria revealed that the genes were closely related to a various number of substrates (e.g. carbohydrates) (222). Therefore it was suggested, that these genes are functionally linked and TBDTs play important roles in nutrient uptake for marine bacteria (222). In our metatranscriptome datasets, the transport profile of *Flavobacteria* and *Gammaproteobacteria* was dominated by TBDT (Figure 2b and d). This is in line with a previous study (222), which revealed that the majority of the TBDT sequences in the Global Ocean Survey (GOS) metagenomic data set

(145) originated from the *Gammaproteobacteria* and *Cytophaga-Flavobacterium-Bacteroides* (CFB) group. Moreover flavobacterial TBDT transcripts were accompanied by a higher level of SusD transcripts (Figure 11d). The latter is a known component of the *Bacteroidetes*-specific starch utilization system that binds the starch backbone and directs hydrolysed oligomers towards a dedicated TBDT receptor for uptake (223). Our results clearly support the metaproteome based hypothesis, that *Flavobacteria* are specialized on complex polymer degradation (140). Furthermore, analysis down to the genus level, which was easily enabled by the high resolution based metatranscriptomic approach, revealed an analogous expression profile for the two dominant members, *Formosa* and *Polaribacter*, (Supplementary Figure 2e-f). Likewise, the gammaproteobacterial SAR92 clade appears to have a similar nutrient strategy (Supplementary Figure 2a). The genome of the SAR92 clade member HTCC2207 exhibits up to 17 TBTT transporter genes (222) supporting the assumption that they may benefit from algae bloom by uptake of complex polysaccharides.

Figure 11: Pfam annotations of genes encoding for TonB-dependent transport systems (TBDT), starch utilization system proteins (SusD), ATP binding cassette (ABC), tripartite ATP independent (TRAP) and tripartite tricarboxylate transporters (TTT). a) *Bacteria*, b) *Gammaproteobacteria*, c) *Alphaproteobacteria* and d) *Flavobacteria*



Unlike the complex polymer degrading bacteria, *Reinekea* exhibited a high expression of ABC and, to a lesser extent, TRAP transporter (Supplementary Figure 2b). *Alphaproteobacteria* also showed high expression rates for monomer transporters such as ABC and TRAP (Figure 11c). The same picture was detected for *Roseobacter* and SAR11 clade (Supplementary Figure 2c-d). This transporter expression profile agrees with previous genome studies of marine microbes (190,222), which revealed ABC transporter as the dominant type in species like *Roseobacter denitrificans* OCh 114 (TBDT: 1 gene; ABC: 110 genes) and *Candidatus* Pelagibacter ubique HTCC1062 (TBDT: 0 gene; ABC: 24 genes).

The main results from the metatranscriptome are in line with the previously analysed metaproteome (140). However, even though expression profiles agreed on the class level, small variances occur on the genus level. Based on the metaproteome analysis (140), *Roseobacter* show a greater expression of transporters than the more abundant SAR11 clade. Our metatranscriptomic data detected the opposite, with up to 16 fold higher amount of membrane transporter transcripts for SAR11 (Supplementary Figure 2c-d). This might be a result of fast mRNA degradation within *Roseobacter* cells. Interestingly, in both spring samples *Roseobacter* feature more 16S cDNA than 16S pyrotags or 16S metagenome reads (Supplementary Figure 1). Previously, Yu et al (224) suggested that high expression levels of rRNA and fast mRNA degradation possibly help bacteria to respond quickly to changing environmental conditions. Keeping in mind that the metaproteome revealed a high number of transporter proteins in comparison to SAR11, which could not be reflected in the metatranscriptome, we suggest that the low amount of detected transcripts are in fact a result of fast mRNA degradation coupled with high rRNA expression. Our results support not only the cellular strategy to an environmental stimulus as described by Yu et al. (224), but also provides another indicator that *Roseobacter* adapt readily to changing nutrient conditions induced by a phytoplankton bloom (200).

Further evidence of taxonomically distinct membrane transporter profiles are provided by additional metatranscriptomic data addressing the cytoplasmic transmembrane components of the TonB complex (ExbB and ExdD) and bacterial extracellular solute-binding proteins (SBP). Expression of ExbB and ExdD is clearly dominated by *Flavobacteria* and *Gammaproteobacteria,* and exhibits a peak in the early algae bloom phase accompanying TBDT expression maxima (Supplementary Figure 3a). On the contrary, SBP encoding genes were almost exclusively expressed by *Alphaproteobacteria* (Supplementary Figure 3b) and in particular of members of the SAR11 clade. SBP are known to be associated with ABC and

TRAP transporters (225-227) binding extracellular solutes for transport across the bacterial cytoplasmic membrane.

Both the metagenome and metatranscriptome revealed a pronounced and taxonomically distinct peak in the abundance of carbohydrate-active enzymes (CAZY) (174). Gene densities of prevalent CAZymes involved in external carbohydrate degradation within the metagenomes was analysed with respect to their maxima at different sampling points. The majority could be assigned to the enzyme class glycoside hydrolases (GH) which allows hydrolysis and/or rearrangement of glycosidic bonds featuring a rich diversity of putative cellulases and hemicellulases (228,229). The metatranscriptome confirmed expression of several CAZymes (Supplementary Figure 4) which would provide an advantage for certain community members, allowing them to benefit from the degradation of complex algae polysaccharides.

For example, GH16 (mainly laminarases) expression was dominated by *Flavobacteria* and *Gammaproteobacteria* (Supplementary Figure 5a). Laminarases are expected to be involved hydrolysis of plant cell walls (230) indicating increased algae cell degradation by bacteria. Our findings are in line with the previous metaproteome data and support the conclusion (140) that during the initial algal die-off phase intact algal heteropolysaccharides became available positively selecting for specialized *Flavobacteria* and some *Gammaproteobacteria*. Additionally, our data revealed, that several transcripts could be assigned to *Formosa*, *Polaribacter* and – to a lesser extent – SAR92 (Supplementary Figure 6a). Likewise, cellulose degrading GH3 exhibits a similar expression profile during the late bloom, although comparatively more transcripts mapped to SAR92 and less to *Polaribacter* (Supplementary Figure 6b).

Unlike the metaproteome, the metatranscriptome revealed GH30 (β-D-fucosidases) expression for both, *Flavobacteria* and – to a lesser extent – *Gammaproteobacteria*. The latter might have been below the detection limit of the proteome analysis.

Transcripts encoding for members of the GH13 (mainly alpha amylases) families mapped to *Flavobacteria* and *Gammaproteobacteria* (Supplementary Figure 5g), thus, reflecting the CAZyme gene abundance in the corresponding metagenome. Several genes could also be mapped to *Reinekea* and *Polaribacter* in the post-bloom phase (Supplementary Figure 6e). GH13 member alpha amylases are known members of the starch utilization system in *Bacteroides thetaiotaomicron* (231,232). Therefore we believe that this enzyme might interact

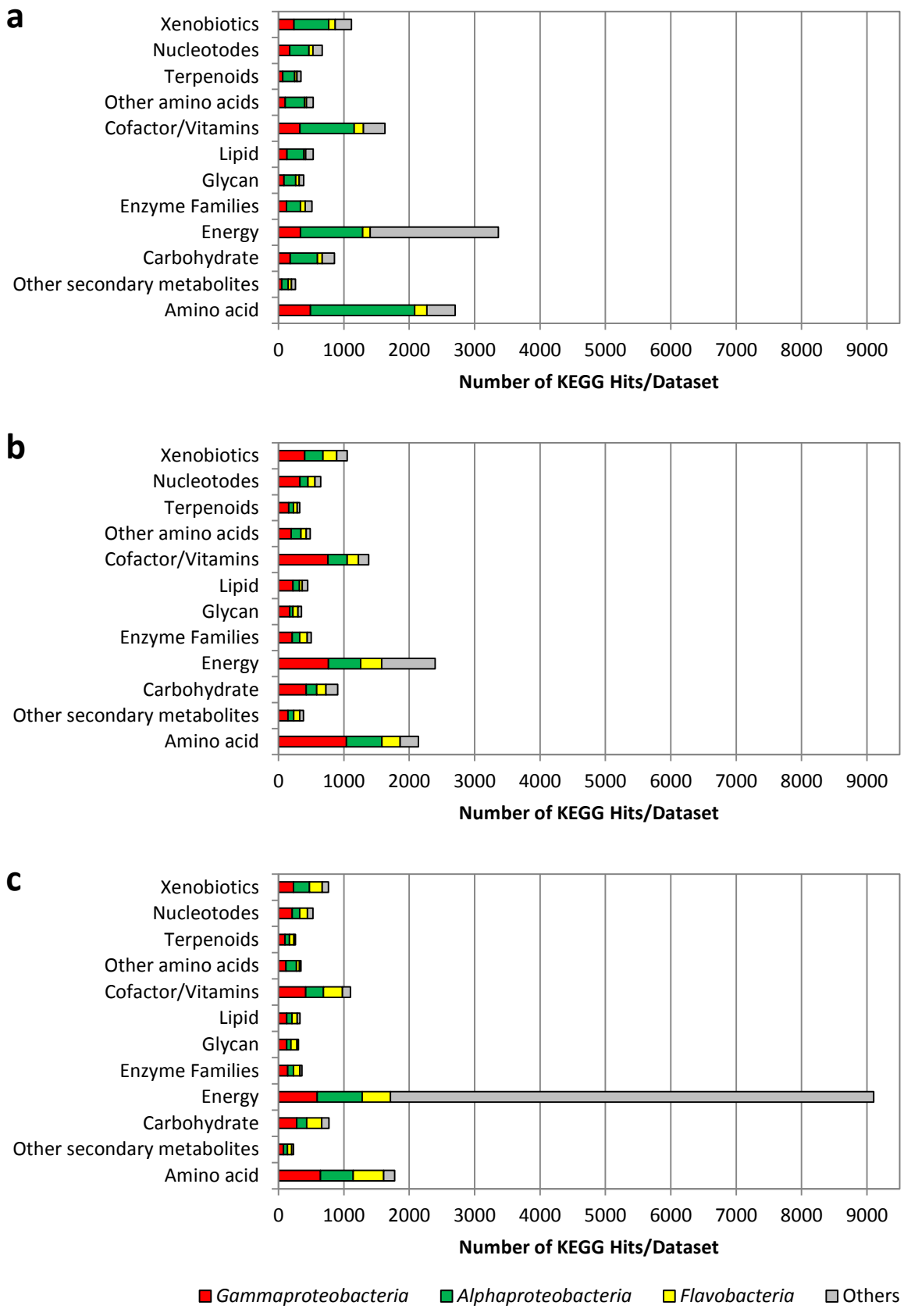
with the SusD and direct hydrolysed starch oligomers towards a dedicated TBDT receptor for uptake in *Flavobacteria*.

Metatranscriptomics also enabled analysis of expressed CAZyme in more detail - a clear advantage over metaproteomics. For instance, CBM50 transcripts feature a taxonomically diverse expression throughout the spring bloom (Supplementary Figure 5c). The carbohydrate-binding-module (CBM) is a defined module within larger enzymes allowing them to bind to carbohydrate such as cellulose. The expressed CBM50 is described as LysM peptidoglycan-binding domain and was originally identified as a component of bacterial lysins (55). Thus the expression most likely reflects increased bacterial cell mortality towards the end of the algae bloom.

Later in the bloom, the CAZyme peak was accompanied by sulfatases (Supplementary Figure 7), which are required for degradation of the many sulphated algal polysaccharides. As seen in the metaproteome, the majority of the transcripts mapped to *Flavobacteria*, confirming their distinct role in algae decomposition. Detailed analysis also confirmed *Polaribacter* expression, although *Formosa* exhibited an even higher amount of sulfatase encoding transcripts. The latter has not been revealed by metaproteomics, thus underlining the higher resolution power of metatranscriptomics even on genus level.

Our study also confirmed expression of the light-dependent Proteorhodopsin (PR) in all three samples (Supplementary Figure 8). However in contrast to a previous study (152), a stable abundance throughout the year without any seasonal fluctuations was not seen. Rather, an increase of the PR encoding transcripts was detected as response to the algae bloom. The first pronounced peak of PR transcripts could be assigned to *Gammaproteobacteria*, of which one third was expressed by members of the SAR92 clade (data not shown), which are known to possess several PR genes (218). Within the *Alphaproteobacteria* class, transcripts were exclusively expressed by SAR11 clade members (data not shown). The expression profile of this species also revealed a pronounced peak of PR towards the end of the phytoplankton bloom (14.04.2009). At that time point the SAR11 abundance is still low, but Teeling et al. proved that the proportions increased after the relative *Flavobacteria* and *Gammaproteobacteria* abundances diminished (140). Therefore we believe that the high PR expression might confer a fitness advantage (49) and allows SAR11 clade members to remain within the community at low levels during the algae bloom, and quickly regain its pre-bloom abundances.

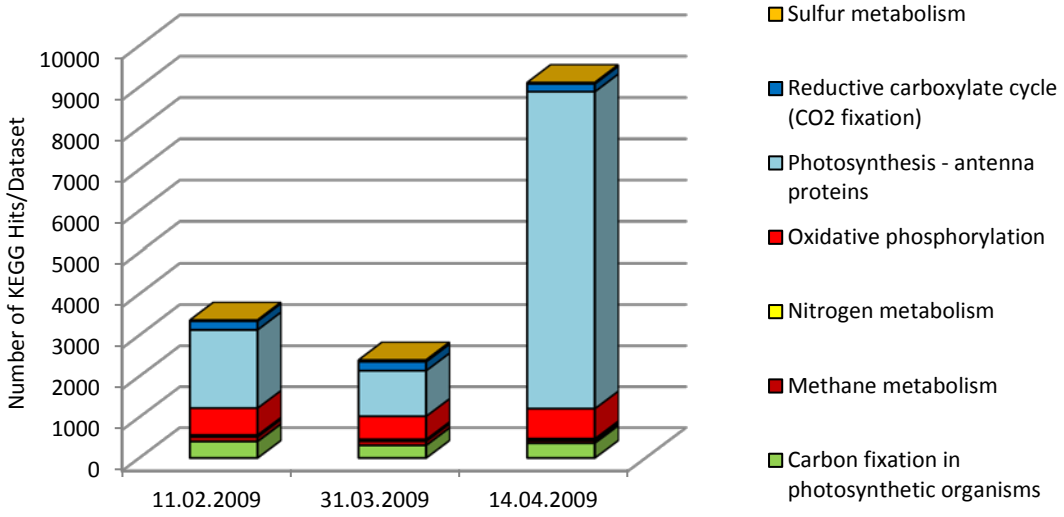
Figure 12: Functional assignment of transcripts based on Kyoto Encyclopedia of Genes and Genomes (KEGG) of selected sampling points a) 11.02.2009, b) 31.03.2009 and c) 14.04.2009).



Metaproteome expression profiles of several phosphate and phosphonate transport systems could only be partially confirmed. No distinct increase or variability of the expression level was detected. Presence of phosphate ABC-type transporters encoding transcripts could only be demonstrated by functional assignment based on the Kyoto Encyclopaedia of Genes and Genomes (KEGG) (216). The majority of those were expressed by *Alphaproteobacteria*, in particular SAR11 clade, supporting previous studies (189). Transcripts encoding for phosphate ABC-type transporters could not be assigned to *Gammaproteobacteria* as indicated by the metaproteome. Moreover, phosphate:sodium symporter expression was rather taxonomically diverse but dominated by *Flavobacteria*. In summary, low transcription levels of phosphate/phosphonate transport systems were detected without any distinct maxima, thus, an increase over the progression of the bloom as described previously could not be verified (140). Although we were able to show a relative constitutive expression of the phosphate uptake regulator PhoU and some transporters the expected increase with regard to phosphate limitation conditions (233) was missing. This discrepancy might arise from posttranscriptional regulation (116), mRNA degradation or detection limits.

In order to expand our knowledge of the metabolic profile of the bacterial community, transcripts were assigned to functional categories based on KEGG with respect to metabolic classes (Figure 12). On all three sampling dates, the majority of sequences could be assigned to amino acids, energy, xenobiotic and cofactors and vitamin metabolism. In winter *Alphaproteobacteria* show a higher metabolic potential which is not surprising due to the higher cellular abundances.

Figure 13: Functional assignment of transcripts based on Kyoto Encyclopedia of Genes and Genomes (KEGG). Detailed view of transcripts assigned to energy metabolism.



Surprising is the striking peak in the energy metabolism towards the end of the algae bloom (14.04.2009) belonging to another group distinct from the dominant three taxa. Detailed analysis revealed that this occurrence was induced by a high number of sequences (up to 7678 KEGG hits per dataset) assigned to 'Photosynthesis – antenna proteins' (Figure 13). Pfam annotations also yielded an enormous increase of transcripts encoding for photosynthetic reaction centers and to a lesser extent photosystems I and II. This is supported by a similar genetic distribution within the metagenome (data not shown). Taxonomic analysis of the transcripts revealed that this striking increase could be assigned to *Cyanobacteria* according to NCBI taxonomy (234). On this sampling day, the German Weather Service (Deutscher Wetterdienst, http://www.dwd.de) measured 10.1 hours of sunshine, which could have induced the transcription of the photosynthetic machinery. This is in line with the study of Liu et al. (80), who demonstrated that *Cyanobacteria* transcription can increase dramatically when fully illuminated. For comparison, the lowest number of 'Photosynthesis – antenna proteins' encoding genes (1116 KEGG hits per dataset) was detected on 31.03.2009 with only 0.2 hours of sunshine. The slightly higher amount (1908 KEGG hits per datset) in winter can be explained by the good weather conditions with regard to 6.7 hours of sunshine. However, based on the analysis of 16S cDNA, 16S pyrotags and 16S metagenome the relative abundance of *Cyanobacteria* is very low (on average below 0.9%) indicating not only low abundances but also low metabolic activity. Although, Poretsky et al. (77) showed that *Cyanobacteria* can dominate the metatranscriptome with a twofold higher representation than in the 16S rDNA diversity studies, the large amount of transcripts is still surprising. We would like to note that 16S rDNA analysis based on the SILVA database (36) revealed on average 9% of reads assigned to chloroplasts when the striking peak of the photosynthetic expression occurred (16S cDNA: 7%; 16S pyrotags: 3%; 16S metagenome: 8%). Unfortunately the NCBI taxonomy lacks distinction between *Cyanobacteria* and chloroplasts as available for ribosomal RNA e.g. in SILVA. Consequently it is worthwhile to speculate that the peak most probably originated from highly active algae on this day.

4.4. Conclusion

Several studies focusing on the gene expression profiles of marine microbial communities have been published recently. However, in most cases research has been restricted to metagenomics in combination with either metaproteomics or metatranscriptomics. To our knowledge, no marine community has been analyzed with a 'full omic' approach using DNA, protein and mRNA data in order to complement each other. Here, we demonstrated a good

correlation of the meta-genome, -transcriptome and -proteome level for the abundant transcripts supporting the previously described substrate controlled bacterial succession (140). High throughput metatranscriptomic data also provided additional information on species level underlining its high resolution power. For example the coherent expression of glycolsyl hydrolases (e.g. GH16 and GH13) and sulfatases most likely allows *Flavobacteria* to decompose complex algae polysaccharides resulting in an increasing availability of sugar oligomers and monomers. The latter is of particular importance for bacteria like *Roseobacter,* whose opportunistic life style allows them to benefit from the changing nutrient conditions. Therefore we support the hypothesis of Teeling et al. (7) that algae substrate availability is a crucial factor for defining a series of ecological niches in which specialized community members could grow.

Simultaneous detection of 16S RNA and mRNA reads revealed a first lead on identifying active members of the community. In particular, *Reinkea* and *Roseobacter* appeared to rapidly respond to environmental changes. High amounts of 16S rRNA copy numbers most likely correlate with the rate at which microbial cells adapt to nutrient availability. Taking into account that metabolically active members contain a higher expression rate of rRNA than starved or inactive cells, 16S cDNA has the potential to serve as screening tool revealing the fitness status of a microbial community.

To conclude, a combined meta- 'omics' approach helps in raising the confidence level of the conclusions, but also sheds light from complementary angles onto the black box of microbial communities responding to environmental stimuli. The integrated interpretation of the diversity data as well as the reconstruction of the dominant metabolic processes and their seasonal changes might create a basis for environmental monitoring in the future.

Supplementary Data

Supplementary Material is available at chapter 9.

Acknowledgments

We acknowledge Jack A. Gilbert (Argonne National Laboratory, Argonne, IL, USA) and Bernhard M. Fuchs (Max Planck Institute, Bremen, Germany) for critical discussion of this work. E. Karamehmedovic and M. Meiners for helping with the laboratory work. G. Gerds and A. Wichels from the Alfred Wegner Institute (Bremerhaven, Germany) for supporting and performing the water sampling. Laura Stusiak, Hannah Marchant and Greta Reintjes for critical reading of the manuscript

Funding

This work was supported by the Max Planck Society and the Federal Ministry of Education and Research Germany [grant number 03F0480D]. Funding for open access charge: Max Planck Society.

5. Chapter

Expression of sulfatases in *Rhodopirellula baltica* and the diversity of sulfatases in the genus *Rhodopirellula*

Authors: Carl-Eric Wegner, Tim Richter-Heitmann, Anna Klindworth, Christine Klockow, Michael Richter, Tilman Achstetter, Jens Harder and Frank Oliver Glöckner

Status: draft to be submitted to Microbial Genomics

Contribution: design, supervision and critical discussion of the experimental part of project, and optimization of experimental procedure in terms of RNA extraction and cDNA synthesis

5.1. Abstract

The whole genome of *Rhodopirellula baltica* SH1^T, published nearly 10 years ago, already revealed a high amount of sulfatase genes. So far, little is known about the diversity and potential functions mediated by sulfatases in *Planctomycetes*. We combined *in vivo* and *in silico* techniques to gain insights into the ecophysiology of planktomycetal sulfatases. Comparative genomics of nine recently sequenced *Rhodopirellula* strains detected 1120 open reading frames annotated as sulfatase (Enzyme Commission number (EC) 3.1.6.*). These were clustered into 173 groups of orthologous and paralogous genes. To analyze functional aspects 709 sulfatase protein sequences from these strains were aligned with 66 sulfatase reference sequences of reviewed functionality. Our analysis yielded 22 major similarity clusters, but only five of these clusters contained *Rhodopirellula* sequences homologous to reference sequences, indicating a surprisingly high diversity. Exemplarily, *R. baltica* SH1^T was grown on different sulfated polysaccharides, chondroitin sulfate, λ-carrageenan and

fucoidan. Subsequent gene expression analyses using whole genome microarrays revealed distinct sulfatase expression profiles based on substrates tested. This might be indicative for a high structural diversity of sulfated polysaccharides as potential substrates. The pattern of sulfatases in individual planctomycete species may reflect ecological niche adaptation.

5.2. Introduction

For a long time, bacterial sulfatases attracted little attention, as the majority of the known bacterial genomes contains only low copy numbers of sulfatase encoding genes [EC 3.6.1.*]. *Rhodopirellula baltica* $SH1^T$ was the first organism sequenced featuring a high number of 110 sulfatases (6). Strain $SH1^T$ is a marine, aerobic and heterotrophic member of the *Planctomycetes.* The pear-like shaped cells divide in a budding-like manner. Adult cells are non-motile, display a polar cell organization and are known to attach to surfaces and to form aggregates, enabled by a holdfast substance, secreted from the vegetative cell pole (235,236).
Ongoing sequencing efforts revealed that this unexpected finding is indeed a characteristic of the *Planctomycetes-Verrucomicrobia-Chlamydia* (PVC) superphylum, i.e. *Lenthisphaera araneosa* (226), *Planctomyces brasiliensis, and Planctomyces maris* feature more than 100 and partially even more than 200 sulfatases (Figure 14).
Sulfatases catalyze the hydrolytic cleavage of sulfate esters and sulfamates. Three distinct classes of sulfatases have been identified so far. Group I sulfatases (formylglycine-dependent sulfatases) are well-known and widely distributed in eukaryotes and prokaryotes. Group II sulfatases (α-ketoglutarate-dependent dioxygenase superfamily alkylsulfatases) and group III sulfatases (Zn^{2+}-dependent alkyl sulfatases) have been recently discovered and only few examples are known (237,238). Substrates range from sulfated proteoglycans and conjugated steroids to smaller aromatic sulfate esters (239).

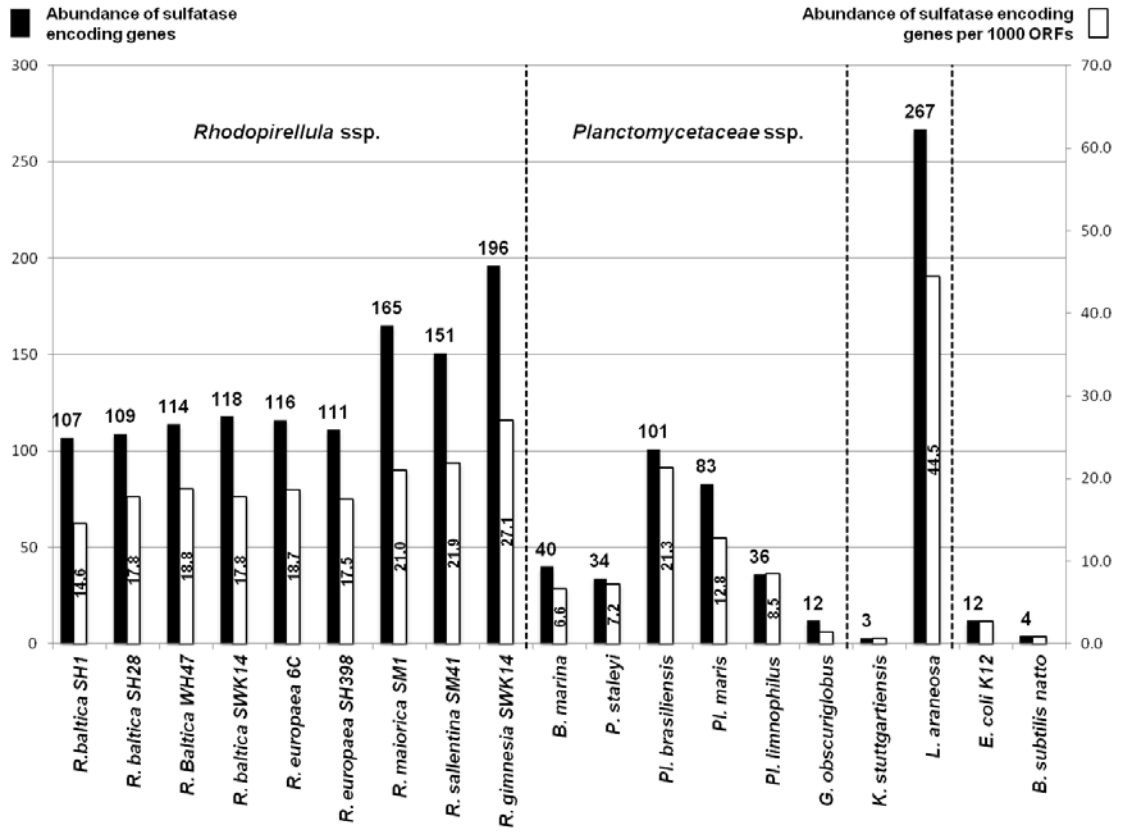


Figure 14: Abundance of sulfatase encoding genes in a number of marine bacteria of the PVC superphylum in comparison to typical strains of the model organisms *E. coli* and *B. subtilis*. The left bars (black) show the absolute amount of genes assigned to this functions, while the right bars (white) give the relative abundance of sulfatase genes per 1000 ORFs. Numbers for the genus of *Rhodopirellula* were obtained by manual assignment of partial sequences to established clusters of homologous genes. Numbers for the other genomes were derived from HMMER3 scans versus the PFAM 25.0 database for the sulfatase model (217), from the UniProt-KB databases, and original publications, respectively. During the process of annotation quality control, the originally stated number of 110 sulfatase encoding genes (6) in the *R. baltica* SH1^T genome was downgraded to 107. Abbreviations: Rba – *Rhodopirellula baltica*; Bmar – *Blastopirellula marina* DSM 3645; Psta – *Pirellula staleyi* DSM 6068; Pbra – *Planctomyces brasiliensis* DSM 3505; Pmar – *Planctomyces maris* DSM 8797; Plim – *Planctomyces limnophilus* DSM 3776; Gobs – *Gemmata obscuriglobus* DSM 5831; Kstu – *Candidatus Kuenenia stuttgartiensis*; Lara – *Lentisphaera araneosa* ATCC BAA-859; Ecoli – *Escherichia coli* K12; Bsub – *Bacillus subtilis* subsp. natto BEST195.

Group I sulfatases share a high structural and sequence similarity. They feature a conserved amino acid signature including a core pentapeptide (C/S-x-P-x-R), followed by (x(4)-T-G), commonly referred to as sulfatase signature sequence I. The cysteine or serine residue within this signature sequence is posttranslationally modified to a catalytically active formylglycine (FGly). Group I is divided into Cys- and Ser-type sulfatases. Ser-type sulfatases were exclusively found in prokaryotes, while the Cys-type has been detected in both eukaryotes and prokaryotes. Two different pathways for the formylglycine formation were discovered.

Formylglycine generating enzymes (FGE) mediate the first mechanism which specifically requires a cysteine residue (240). The second system involves anaerobic sulfatase modifying enzymes (anSME) which are able to convert cysteine or serine in the active site (241). *Escherichia coli* mutants carrying gene deletions in both described maturation systems still expressed functional sulfatases. Therefore, a third, uncharacterized maturation system seems to exist (242). The currently favored mechanism of sulfatase catalysis is a transesterification mechanism, utilizing the hydratation of the formylglycine to a geminal diol. In the course of two subsequent nucleophilic attacks, the organic moiety and the sulfate group are released from the initial substrate (Figure 15) (243,244).

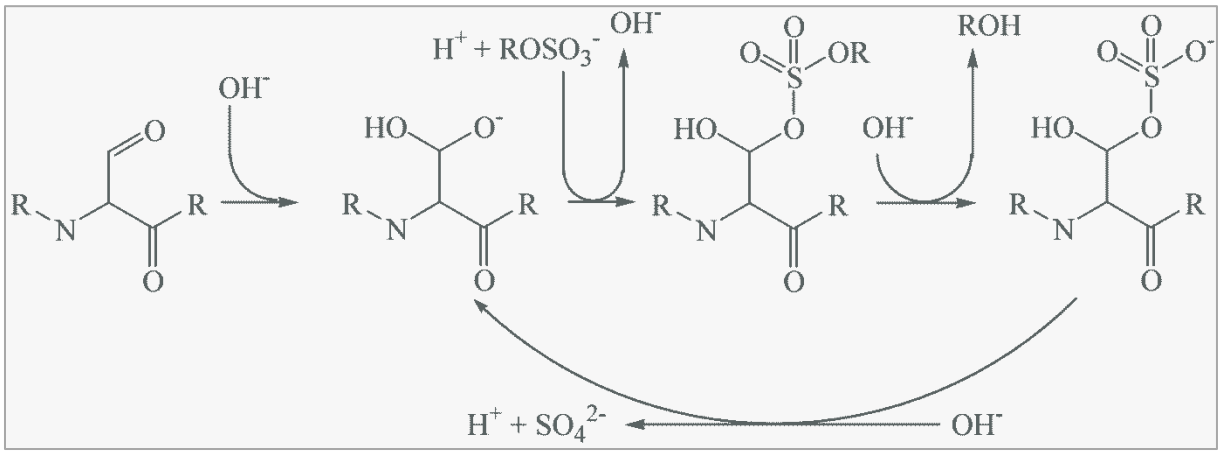


Figure 15: The proposed transesterification mechanism of group I. sulfatases. The hydrated formylglycine residue, a geminal diol functions as nucleophile. In the course of two nucleophilic attacks the organic rest and the sulfate are released.

It has been suggested that the high number of sulfatases found in *Planctomycetes* could play a major role in the degradation of sulfated polysaccharides in their environment. Indeed, the degradation of sulfated biopolymers seems to be a prominent part of their physiology (245,246). Organisms related to *R. baltica* SH1^T were found to be associated with macroalgae in Portuguese coastal waters (247) and the dominating lineage in biofilms on kelps (248). Algal cell walls are known to contain plenty of sulfated carbohydrates, such as ulvan or fucoidan (249,250). Another study suggested that *R. baltica* SH1^T is able to convert partially sulfated algal carbohydrates such as carrageenans (251). These findings support the hypothesis that *R. baltica* SH1^T might be specialized in degrading sulfated polysaccharides in its natural habitat.

Further, transcriptome studies with this model organism demonstrated that also in the absence of any sulfated substrate, 11 sulfatase genes are up- or down-regulated in response to different stresses (126). The same authors additionally investigated transcriptome-wide gene

expression changes at different stages of the life cycle (252). 12 sulfatases were found to be differentially expressed. These results suggest a currently unknown role of sulfated molecules and their hydrolysates in the cellular physiology of *R. baltica* SH1^T.

In this study, we phylogenetically assessed the diversity of sulfatase genes of *R. baltica* SH1^T, together with sulfatase genes found in eight permanent draft genomes of strains representing five species including *R. baltica* SH1^T, which were obtained in a study covering the biogeography of planctomycetes in European seas (Table 3) (253,254). Growth experiments on a diverse set of sulfated polysaccharides were conducted with whole genome gene expression profiles to identify the substrate specificity and eventually the cooperation of multiple sulfatases involved in the degradation of sulfated polysaccharides.

Table 3: List of analyzed *Rhodopirellula* genomes, in addition to the type strain *Rhodopirellula* baltica SH1^T. 16S rDNA similarity values were calculated against the reference type strain. The average nucleotide identity (ANI) between the type strain genome and 8 draft genome sequences was determined by using the *in silico* DNA-DNA hybridization method of the JSpecies (255) software with default parameters. Operation taxonomic unit (OTU) classification is referring to the original clustering as suggested by Winkelmann et al. (254).

Strain	OTU[1]	Sample Site[2]	16S rDNA similarity[2]	ANI	Proposed name[3]	Genome size[mb]	Predicted ORF	ACC (GenBank)
WH47	A	Sylt, Germany	>99.6	97.35	*R. baltica*	6.24	6059	AFAR00000000
SH28	A	Kiel Fjord, Germany	>99.6	97.05	*R. baltica*	6.38	6140	
SWK14	A	Tjärnö, Sweden	>99.6	97.25	*R. baltica*	6.59	6633	
6C	B	Porto Cesareo, Italy	>99.5	88.38	*R. europaea*	6.42	6210	
SH398	B	Kiel Fjord, Germany	>99.5	88.48	*R. europaea*	6.63	6361	
SWK7	F	Tjärnö, Sweden	98.6	70.42	*R. gimnesia*	8.78	7242	
SM1	D	Pt. Andratx (Mallorca), Spain	96.1	68.73	*R. maiorica*	8.88	7847	
SM41	C	San Cataldo, Italy	97.7 - 97.9	70.47	*R. sallentina*	8.19	6889	

[1] Winkelmann *et al.* (254) [2] Winkelmann *et al.* (253)

5.3. Material and Methods

Bioinformatic analysis

Genomic data

The procedures used to obtain the permanent draft genomes are described in detail in Glöckner et al. (6) and Frank (256) , respectively.

Sulfatase gene identification

Sulfatase encoding genes were identified with HMMer3 scans versus the PFAM database 25.0 with an E-value threshold set to 1.0E-05. This procedure was complemented by performing an automatic annotation with MicHanThi (www.megx.net/michanthi/michanthi.html), to avoid missing genes incorrectly not being identified with HMMer3.

Identification of orthologous and paralogous genes

Full gene sequences were analyzed with OrthoMCL 2.0 (257) using default parameters, which combines reciprocal best match (RBM) BLAST and Markov clustering to identify paralogous and orthologous gene families. Partial sequences were aligned to obtained clusters of paralogous and orthologous groups with the BLASTP alignment algorithm. A threshold of 50% position identities to at least one member of a best matching cluster was used for cluster assignment. Thus, sequences representing a single gene, but being scattered between several contigs, could be identified.

Phylogenetic tree construction

Overall, 709 sulfatase sequences of *Rhodopirellula* species were selected for phylogenetic analysis. Redundant sequences from strains of the same species were removed from the final data set to save calculation time. A set of 66 reviewed sulfatase sequences of known substrate specificity from a variety of species were retrieved from the UniProt database and aligned to the *Rhodopirellula* gene set, in order to gain functional information on the unknown proteins. MAFFT (FFT-NS-I; (258)) was applied for the alignment of the final dataset of 765 sequences in Jalview 2.6.1 (259). Maximum Likelihood phylogeny was carried out with RAxML 7.2.8 (260), which was executed on the Teragrid server of the Cipres Science Gateway (261). For the evolutionary model, the heuristic CAT approximation with the JTT

substitution matrix was chosen. RAxML was called with the command line - raxmlHPC-HYBRID-7.2.8 -T 6 –f a –m protcatjtt –N 100 –x 12345.
100 replicates (bootstraps) were calculated, with the confidence cutoff being set to 50 for each node in the consensus tree. The obtained tree was visualized with Archaeopteryx 0.957 (262). Active site conservation was checked with Weblogo 3.0 (263).

Transcriptome-wide gene expression analyses

Cultivation of *R. baltica* SH1^{T}

R. baltica SH1^{T} was cultivated in minimum mineral medium (MMM) supplemented with individual sulfated carbohydrates as carbon source (supplementary material). Glucose has been set as reference carbon source. Fucoidan (GlycoMix, Reading, UK, product ID: PSA10), λ-carrageenan (Sigma-Aldrich, Munich, Germany, 22049) and chondroitin sulfate (Sigma-Aldrich, C4384) have been chosen as substrates of interest. Pre-cultures for high-volume cultures (500 mL) were set up by inoculating small-volume cultures (50 mL) of MMM supplemented with glucose. After two transfers, the volume of the pre-cultures was stepwise increased by 50 mL MMM. The final volume of pre-cultures was 150 mL. The growth of cultures was monitored by regularly measuring the $OD_{600\ nm}$. As soon as mid-exponential phase was reached, at an OD of 0.6 to 0.9, high-volume cultures were inoculated with 75 mL of pre-culture (15% v/v). As negative control, one high volume culture was set up with medium without being supplemented with any substrate. Cultures were incubated at 28 °C under shaking using baffled Erlenmeyer flasks until mid-exponential phase (OD 0.6-0.9) was reached (incubator: INE 800, Memmert, Schwabach, Germany; shaker: KS501, IKA Labortechnik, Staufen, Germany).

Determination of basic growth parameters

Starting from two pre-cultures (50 mL) which had been transferred twice after been grown to mid exponential phase on glucose, three cultures (50 mL) per substrate of interest (chondroitin sulfate, λ-carrageenan, fucoidan and glucose as reference) were prepared with a 10% (v/v) inoculum (5 mL). The initial OD_{600nm} was determined and monitored over one week. As negative control, three cultures had no substrate. As positive control, three cultures were grown on the complex medium M13a supplemented with casamino acids (German collection of microorganisms and cell cultures, Pirellula medium 600a, (130). Growth curves allowed the calculation of growth rates and doubling times.

Extraction, clean up and quality assessment of total RNA

Cell material for downstream processing was harvested by centrifugation and was kept at -20°C (-80 °C for long term storage) until being processed. Stored cell pellets were thoroughly resuspended in 1-3 mL of TRI reagent (Applied Biosystems, Darmstadt, Germany). The suspension was incubated for 5 min at room temperature. Cells were lysed by beadbeating (lysing matrix B, material: 0.1 mm silica spheres; MPBiomedicals, Berlin, Germany) applying a FastPrep 24 automated homogenizer (MPBiomedicals). Three steps of 30 sec (speed: 6 m/s) were performed, while cooling the tubes on ice between beadbeating steps. After the third step, the beadbeater tubes were incubated on ice for additional 10 min. Next, beadbeater tubes were centrifugated at 4 °C for 10 min (5415 C, Eppendorf, Hamburg, Germany; 13200 rpm, rotor: FA-45-24-11). Supernatants were transferred into RNase-free, sterile 1.5 mL Eppendorf cups. 200 µL of ice-cold chloroform was added per sample. Suspensions were thoroughly mixed by vortexing for 20 sec, followed by a 10 min incubation step at RT. A further centrifugation step was carried out (4 °C, 15 min, 13200 rpm). The aqueous, upper phase was transferred into new, RNase-free and sterile Eppendorf cups. 1 mL of 100% isopropanol was added, followed by incubation at – 20 °C for 1 hour (hr). After the incubation, a 30 min centrifugation step was performed (4°C, 13200 rpm). The supernatants were discarded and pellets were washed twiced in 75% ethanol. Dried pellets were dissolved in 50-100 µl RNase-free water. Extracted RNA was cleaned using the RNeasy MinElute clean-up kit (Qiagen, Hilden, Germany) following the manufacturer's instructions. The concentration and quality of eluted RNA was determined using a NanoDrop® spectrophotometer (Thermo Scientific, Wilmington, USA). Amount and quality of extracted and cleaned up RNA was also documented by RNA agarose gelectrophoresis.

Single-stranded cDNA synthesis and cDNA labelling

For synthesizing single-stranded cDNA, the SuperScript Direct cDNA labelling Core kit (Applied Biosystems) was used applying random hexamers and following the manufacturer's instructions. Per synthesis reaction/sample, 5-10 µg of extracted RNA were utilized in the three hour reverse transcription step at 46 °C. The reaction was halted by incubating at 95 °C for 5 min. Samples were hydrolyzed by adding 15 µL of 0.1 M NaOH, being incubated at 65 °C for 15 min and adding 15 µL of 0.1 M HCl. Single stranded cDNA was purified using the QIAEX II Gel Extraction Kit (Qiagen, Hilden, Germany) according to the instructions described in the manual. The concentration of synthesized cDNA was determined using a

NanoDrop® spectrophotometer (Thermo Scientific) using nuclease-free water as blank. Besides, amount and quality of synthesized and purified cDNA was cross-validated by DNA agarose gelectrophoresis. Samples have been directly labeled applying the Platimum *Bright*^{Tm} Alexa 546 and Alexa 647 labeling kits (Kreatech, Amsterdam, Netherlands) nucleic acid labeling kits according to the manufacturer`s protocol. Alexa 546 was generally used for glucose reference samples, while Alexa 647 was applied to samples linked to substrates of interest.

Whole genome array hybridization

Detailed information relating to the applied whole genome array of *R. baltica* SH1^{T} and its production is available through the Gene Expression Omnibus database (http://www.ncbi.nlm.nih.gov/geo/) (GEO ID: GPL7654) and from two previous studies (126,252). In brief, the hybridization reaction including denaturing, hybridization, washing and N_2 drying was conducted using a HS 400 Pro hybridization station and respective software (Tecan, Crailsheim, Germany). Arrays have been blocked by pre-hybridization buffer made up by 250 mM NaCl, 5 mM Tris/HCl (pH 8), 50% formamide, 0.5 SSC, 0.05% BSA and 1% blocking reagent (Roche Diagnostics, Mannheim, Germany) for 45 min at 52 °C. Per hybridization reaction, 2 µg of Alexa 546 labelled total cDNA and 2 µg of Alexa 647 labeled total cDNA were pooled and subsequently taken up in a final volume of 100 µl DIG Easy Hyb hybridization solution (Roche Diagnostics, Mannheim, Germany). After blocking the arrays, sample solutions were applied to the arrays, followed by denaturation at 95 °C for 3 min and hybridization at stringent conditions for more than 12 hrs at 52 °C. ULTRArray Low stringency wash buffer (Applied Biosystems) was used for washing slides after hybridization was finished followed by drying the slides using plain N_2. Per comparative analysis, three arrays have been investigated in parallel, using samples originating from biological replicates.

Signal detection, data processing, and analysis

Slides were pre-scanned at a resolution of 50 µm followed by a scan at 5 µm applying a ScanArray Express Microarray scanner (Perkin Elmer, Wellesley, USA). Associated software, ScanArray Express Version 4.0 was used for automatic spot detection and signal quantification referring to both applied dyes. Data quality was enhanced by manually curating spots classified and assigned by ScanArray Express software. Data deduced from image data was further processed using the microarray data analysis software tool MADA

(http://www.mpi-bremen.de/en/MADA.html). Spot intensities were corrected for local background, meaning the spot intensity minus the mean spot background intensity. Signals were assumed to be positive if the mean spot intensity was higher than the mean local background intensity plus twice the standard deviation of the local background intensity. Because each gene is spotted three times per microarray, MADA also compares the quality of the spots among each other with its outlier test. In order to remove poor quality spots from the data sets, standard deviations relating to each spot triplicate are calculated. Subsequently, calculating the deviations is repeated, this time leaving one replicate out. In case that the *de novo* calculated deviation differs more than 50% from the previous, the left out replicate is considered as an outlier. The outlier test is repeated for each replicate. Expression was defined by the ratio and intensity, with R being the ratio (R = log2 (result of channel 2 (sample) / result of channel 1 (control)) and I being the intensity (I = log10 (result of channel 2 (sample) × result of channel 1 (control)). In order to normalize the data, an R versus I plot was done regarding a self-hybridization of reference samples. The reference (*R. baltica* SH1^T grown on glucose) was labelled twice, once with Alexa 546 and once with Alexa 647. Normalization was conducted by LOWESS normalization using a smoothing factor of 0.5. Since at least two hybridizations were done per experiment, expression data from replicates were combined to one expression data point by averaging. A valid expression was assumed if the standard deviation was below 25%. The variability of the self-self hybridization was used as basis for determining the background noise. Differentially expressed genes were determined by setting fixed thresholds taking the background noise of the self-hybridization into account. MayDay (264) was used for analysis of expression patterns in individual data sets. Microarray data was deposited at Gene Expression Omnibus database, GEO ID: GSE35832.

5.4. Results and Discussion

Sulfatase genes in *Rhodopirellula* genomes

In total, 1222 sequences annotated as sulfatases were found in the complete dataset consisting of the recently sequenced draft genomes of eight *Rhodopirellula* strains and the manually curated genome of the *R. baltica* SH1^T type strain. After the correct allocation of partial sequences scattered between different contigs, we could assign 1120 sequences to 173 clusters of ortho- and paralogy, with the latter being a rare exception (Figure 16A). A total of 67 genes appeared to not having close relatives, and are thus considered to represent potential unique substrate specificities.

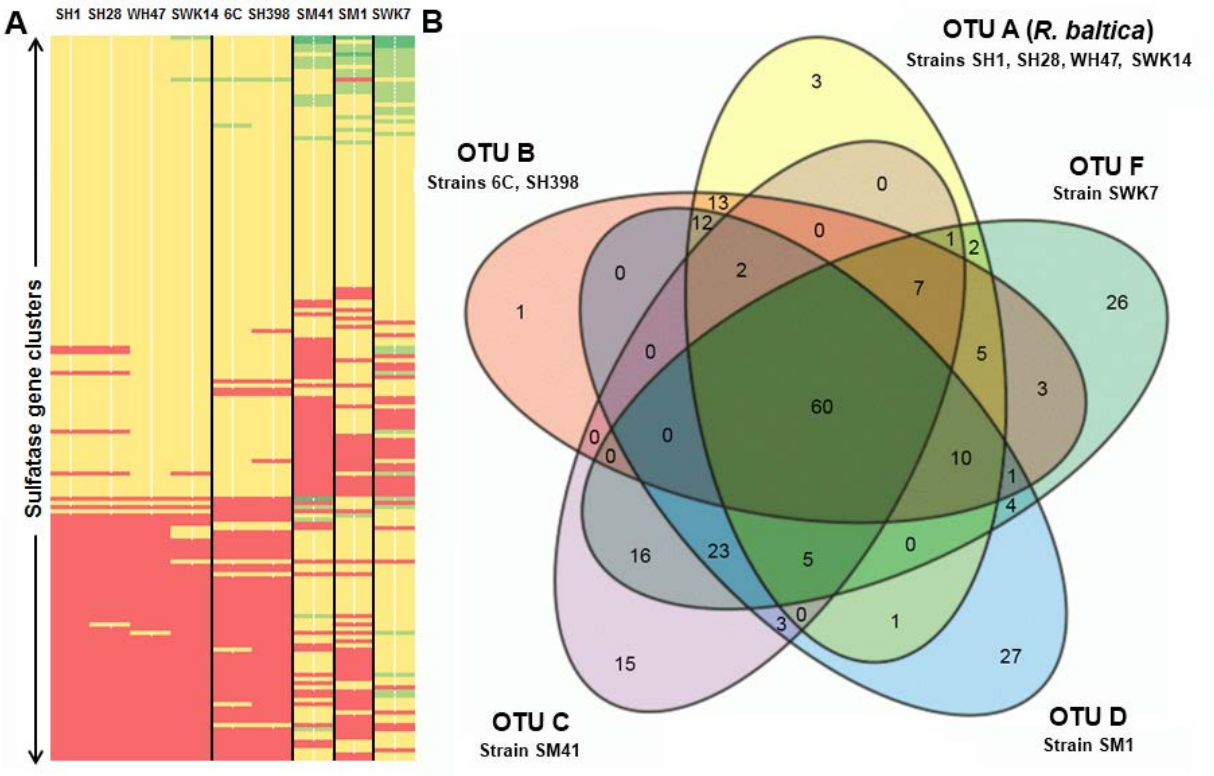


Figure 16: Clusters of ortho- and paralogous sulfatase encoding genes between *Rhodopirellula* strains obtained by OrthoMCL and manual sequence assignment. (A) Conditionally formatted heat map of ortho-/paralogous gene clusters. Red boxes indicate absent genes, while other colors represent varying numbers of observed gene copies (yellow = 1, light green = 2, dark green = 3 copies). (B) A five armed VENN diagram of sulfatase gene distribution between five *Rhodopirellula* species. Data was normalized in a way that paralogous genes were counted as a single hit for the respective species. Genes that were present in at least one strain of a species were counted as a hit for the whole species.

The genus wide 'pangenome of sulfatases' was therefore calculated to 240 singular specimen. A core set of 60 sulfatases occurring in all nine investigated organisms was identified (Figure 16B). Huge intersections were observed for the strains of OTUs A and B. (99 genes were present in at least one strain of both species), and for *OTU C and OTU F* (112 shared sulfatases). *OTU D* features significantly more shared sulfatases with *OTU A* strains (12) than with *OTU B* (zero). Generally, species with higher abundance of sulfatases yielded more unique sequences than paralogs. The close relationship between *OTUs A and B* was also confirmed by phylogenies based on 16S rRNA genes and multi locus sequence analysis (not shown).

Bioinformatic assessment lead to the finding that the vast majority of sulfatase genes in the data set represents single copy genes in their respective genomes. This suggests an immensely diverse range of application for the encoded proteins. Sulfatases being identified as involved

in cellular mechanisms apart from carbohydrate degradation in previous studies (126,252) were in any case conserved in at least three species.

Phylogenetic analysis of sulfatases and active site conservation

Phylogenetic analysis on the protein sequence level was carried out with both Neighbor Joining and Maximum Likelihood methods in order to reveal evolutionary relations and functional capabilities. 709 *Rhodopirellula* sp. sulfatase sequences representing one gene per species and cluster were selected and aligned to 66 sequences of reviewed sulfatases from UniProt, resulting in an alignment with 6429 positions. The sequence lengths varied between 264 and 1829 amino acids (the latter one being a fusion enzyme with two sulfatase domains and an additional domain of unknown function (DUF1680) exclusively found in the genome of *OTU C*. The vast majority of all sequences ranged between 450 and 550 residues in length. Both obtained trees showed the same topology. Figure 17 depicts the Maximum Likelihood tree as unrooted and circular. The early stages of sulfatase evolution showed low confidence values in general.

22 distinct branches with at least two clustered sequences were detected in the tree with three additional single *Rhodopirellula* sp. sequences being unclustered and possibly representing distinct functionality. 19 branches contained sequences of *Rhodopirellula* origin, while the remaining three branches were consisting of reference sequences only: Glucosamine (N-acetyl)-6-sulfatase (GNS), mammalian sulfatases 1 and 2, two *Chlostridium* sulfatases (SULF_CLOP1 and SULF_CLOPE), sulfatase yidJ from *E. coli*, and eukaryotic arylsulfatases arsK were not clustered to any *Rhodopirellula* sequence, respectively. The only known choline sulfatase (betC of *Rhizobium meliloti*) surprisingly did also not group into a cluster, although this annotation is often found in public databases. Five of the major branches contained both known and *Rhodopirellula* sequences (Clusters G, H, I, M, and N, respectively; Table 3), leaving 14 clusters of just *Rhodopirellula* spec. genes, which are not closely related to any sulfatase sequence with known activity. This finding – although not usable for actual functionality prediction without experimental proof – showed an amazing diversity at the sequence level. The diversity is further highlighted by the fact that the well-studied mammalian arylsulfatases are clustering very closely to each other in just three different of the major sulfatase groups in the tree. Unfortunately, at the time the experiment was conducted, no known sequence was obtainable for the class of mucin-desulfating sulfatases, which is also a frequent annotation in bacterial genomes.

We have also been interested in the degree of conservation of the sulfatase signature sequence I of this enzyme class within the major clusters of predicted similar functionality. Cluster O was the only group of sulfatases in this study not featuring a fully developed sulfatase sequence I motif. Consistent with previous findings (265), no Ser-type sulfatase sequence was found within the *Rhodopirellula* dataset. The presence of only Cystein type I sulfatases and the correspondent aerobe FGE maturation system in any genome might reflect the strict aerobic lifestyle of this genus.

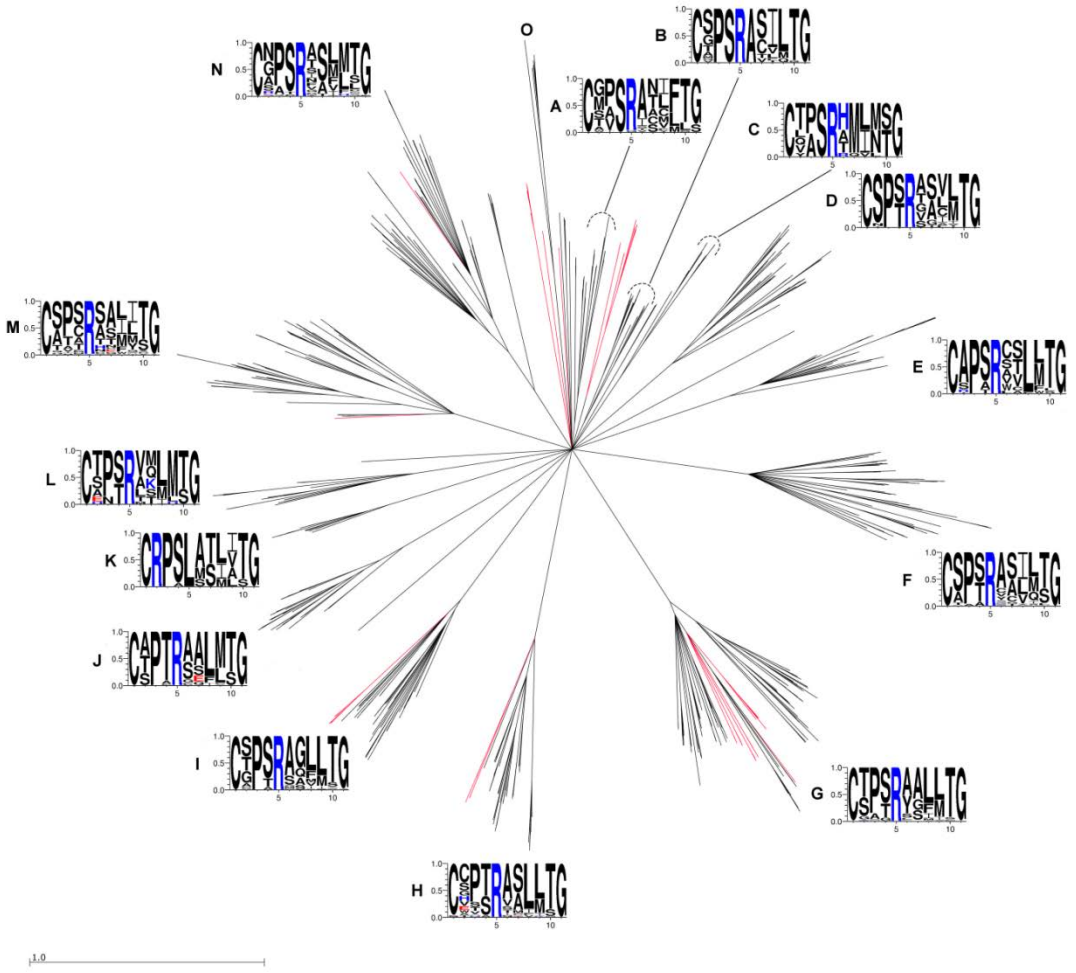


Figure 17: Phylogenetic analysis by Maximum Likelihood method for a set of 775 sulfatase sequences. A circular, unrooted topology is shown. Branches with bootstrap values below 50 were collapsed. For the evolutionary model, the heuristic CAT approximation with the JTT substitution matrix was used. 100 bootstraps were performed. The scale bar corresponds to a genetic distance of one substitution per 100 positions. Red branches represent reviewed sulfatase sequences obtained by the UniProt database. Major branches are named alphabetically in clockwise rotation. The sequence logo depicts the site conservation of the sulfatase signature sequence I as a percentage distribution per site (obtained with WebLogo 3.0 (263)).

Table 4: Overview of major similarity clusters containing both, *Rhodopirellula* spec. and known sulfatase sequences from the UniProt database, their respective positions in the phylogenetic tree as shown in Figure 17, and their function as given in the PFAM and UniProt database. Please note that a reviewed sequence status in UniProt does not necessarily require knowledge of on substrate specificity level.

Reference sequences (UniProt Accessions)	***Rhodopirellula* spec. sequences Tree Cluster ID (Fig. 6)**	**Function**
atsA_Klepn (Q9X759) **atsA_Kleae** (P20713**)** **ars_Pseae** (P51691) **YHJ2_SCHPO** (Q9C0V7)	53 Sequences in 12 homology clusters Cluster H	**Arylsulfatases** of bacterial or yeast origin with unknown substrate specificity.
arsB_Human (P15848**)** **arsI_Human** (Q5FYB1) **arsJ_Human** (Q5FYB0) **(+ other mammalian homologs)**	56 sequences in 17 homology clusters Cluster I	**arsB** (EC 3.1.6.12): Hydrolysis of the 4-sulfate groups of the N-acetyl-D-galactosamine 4-sulfate units of chondroitin sulfate and dermatan sulfate. **arsI/arsJ**: Unknown *in vivo* function.
SPHM_Human (P51688)	85 sequences in 19 homology clusters Cluster M	**N-sulphoglucosamine sulphohydrolase**: Lysosomal hydrolyzation of N-sulfo-D-glucosamine into glucosamine and sulfate.
IDS_Mouse (Q08890) **IDS_Human** (P22304)	90 sequences in 21 homology clusters Cluster N	**Iduronate-2-sulfatase** (EC 3.1.6.13**):** Lysosomal hydrolyzation of 2-sulfate groups from iduronic acids in dermatan sulfate and heparan sulfate.
GALNS_Human (P34059) **arsA_Human** (P15289) **arsD_Human** (P51689) **arsE_Human** (P51690**)** **arsF_Human** (P54793) **arsG_Human** (Q96EG1) **arsH_Human** (Q5FYA8) **(+ other mammalian homologs)** **sts_Human** (P08842) **ars_Hempu** (P14000) **ars_Strpu** (P50473) **aslA_Ecoli (**P25549)	136 sequences in 27 homology clusters Cluster G	**GALNS**[1] (EC 3.1.6.4): Hydrolysis of the 6-sulfate groups of the N-acetyl-D-galactosamine 6-sulfate units of chondroitin sulfate and of the D-galactose 6-sulfate units of keratan sulfate. **arsA (**EC 3.1.6.8): Hydrolysis of cerebroside sulfate. **Mammalian arylsulfatases D,E;F,G, and H** : Unknown *in vivo* function. **sts**[2] (EC 3.1.6.2): Conversion of sulfated steroid precursors. Sea Urchin **arylsulfatases** (EC 3.1.6.1): Unknown *in vivo* function. Bacterial **arylsulfatase** (EC 3.1.6.1): Unknown *in vivo* function

[1] GALNS = N-acetylgalactosamine-6-sulfatase
[2] sts = Sterylsulfatase

From the results, we can report a high conservation for the cysteine (position 1) and the arginine (position 5), within the signature sequence. The proline in position 3 was also strongly conserved in clusters B, D, E, I, J, and K, respectively. The other clusters showed a higher diversity at this position. Strikingly, sequences in cluster K were exhibiting a leucine in position 5, instead of the usual arginine. This transition is ought to have a tremendous effect on the active site configuration, as leucine lacks the positive charge and is significantly smaller. This particular arginine is thought to stabilize the diol moiety of the formylglycine via a hydrogen bridge formed by a secondary amino group (244). Strong diversity inside homology clusters was observed for the other positions of the signature sequence, although every sequence ended with glycine. In summary, a small but observable effect of the active site conservation on the tree topology was found. One can also assume that evolutionary pressure is more likely to be driven by functional conservation than by species separation.

We also scanned all full sulfatase sequences for the occurrence of signal peptides and transmembrane helices with SignalP 4.0 (266) and TMHMM 2.0 (267), respectively. However, the results were found to be inconsistent within members of conserved homology clusters, which suggest problems of common models with compartiments in *Planctomycetes*. Only ten sequences yielded significant signals with four or more predicted helices. At any rate, membrane bound sulfatases were rarely found in the genus *Rhodopirellula*.

Sulfated polysaccharides as growth substrates for *R. baltica* SH1^{T}

As the computational assessment of the sulfatase dataset promised an unexpectedly high diversity in substrate recognition, we tested expression patterns for the model organism *R. baltica* SH1^{T} to challenge this hypothesis. Growing *R. baltica* SH1^{T} on different sulfated substrates revealed varying growth efficiencies. Compared to glucose as reference substrate, the utilization of chondroitin sulfate resulted in higher growth rates (Figure 18). Results from λ-carrageenan were comparable to those from glucose. In comparison to the non-supplemented negative control, no growth was observed for fucoidan. Decreased growth rates and longer doubling times were found for all substrates tested compared to the positive control grown on complex medium.

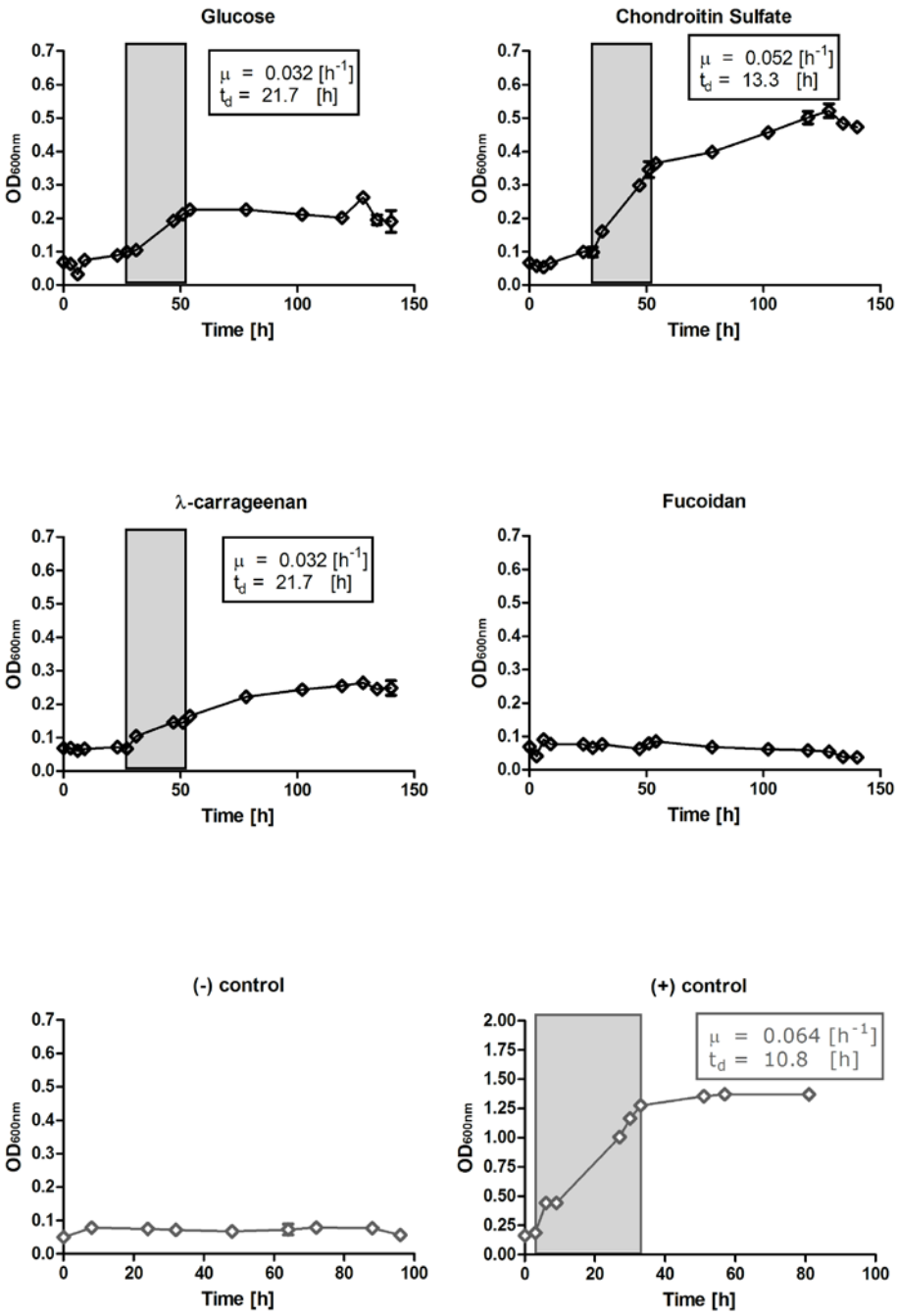


Figure 18: Determination of basic growth parameters relating to *R. baltica* SH1^T cultures grown on different sulfated polysaccharides. Parameters have been determined based on three parallels and for the calculations the indicated time intervals have been taken into account. Average values are given and standard deviations are indicated by error bars. Glucose has been examined as reference substrate. As negative control, three cultures have been set up with medium not containing any substrate. *R.baltica* SH1^T grown on complex medium (M13a + casamino acid) functioned as positive control. μ = growth rate, t_d = doubling time.

Table 5: Expressed and regulated sulfatases in *R. baltica* SH1^T cultures grown on different sulfated polysaccharides. ─ = sulfatase gene was not expressed, + = sulfatase gene was expressed, * = sulfatase gene was upregulated. SignalP (Bendtsen et al., 2004) and THMM (Krogh et al., 2001) webservices were used for determining the presence of signal peptides and transmembrane helices, respectively. E = signal peptide present, T = transmembrane helices present. Numbers indicate the number of gene copies in the respective species.

Gene ID	Phylogenetic Cluster	SigP	TMHMM	Glucose	Chondroitin sulfate	λ-carrageenan	Fucoidan	OTU B (2 strains)	OTU C	OTU D	OTU F
RB406	A	─	─	─	+	─	+	1	3	1	3
RB200	B	**E**	**T**	+	─	+	─	1	1	2	2
RB3403	D	**E**	**T**	+	─	+	─	1	1	1	0
RB4787	D	**E**	─	─	─	+	─	1	0	1	1
RB4815	D	**E**	─					0	0	1	0
RB1477	E	─	─	─	+	─	─	1	2	2	3
RB5146	G	**E**	─	─	+	─	─	1	0	1	0
RB7875	G	**E**	─	+	+	+	+	1	0	1	0
RB13148	G	**E**	─	─	+	─	─	1	3	3	3
RB2367	H	**E**	─	─	+	─	─	2	1	0	1
RB348	I	─	─	─	+	─	+	1	1	1	0
RB3849	I	─	**T**	+	+	+	+	1	0	0	1
RB198	J	**E**	─	+	─	─	+	1	1	1	1
RB9091	J	─	─	+	+*	+	─	1	1	1	1
RB9755	J	─	─	─	+	─	─	0	0	0	1
RB5305	L	**E**	─	─	+	─	─	1	1	1	1
RB3177	M	─	─	─	+	─	─	1	1	1	1
RB5294	M	**E**	─	─	+	+	─	1	2	1	1
RB9549	N	**E**	─	+	─	─	─	1	2	1	1

The comparable or even better growth performance regarding λ-carrageenan and chondroitin sulfate given equal concentrations of substrate applied is probably a consequence of those substrates matching the natural environment of *R. baltica* SH1^T more than glucose. Both, chondroitin sulfate and λ-carrageenan occur in significant amounts in marine environments and also niches inhabited by *R. baltica* SH1^T (268,269). The finding, that *R. baltica* SH1^T is not growing on fucoidan was surprising. Closely related species of *R. baltica* SH1^T are known to dominate biofilms on the brown algae *Laminaria hyperborea*. These brown algae are known to secrete significant amounts of fucoidans. *R. baltica* SH1^T is featuring only a single gene encoding for an α-L-fucoidase. Two other species of this genus (OTUs C and F) were found to bear more than 20 copies of this gene (not shown). Therefore, other species of this genus probably inhabit these ecosystems. In the past, it was proposed that secreted fucoidans can probably function as growth substrate for present marine *Planctomycetes*. However, fucoidans from different algal species can strongly differ in their structure (270,271). In this study fucoidan from *Fucus vesiculosus* was used as growth substrate. The lack of growth during the study is probably due to structural differences between fucoidans of different origin or due to the aforementioned lack of suitable hydrolase activities.

mRNA expression of sulfatases

Differently sized data sets were obtained from microarray analyses. Generally, 1000 to 1500 genes were found to be expressed, representing 14 to 20% of all genes present in the genome of *R. baltica* SH1^T. The fucoidan-related data set was an outlier with only 524 genes. In the context of chondroitin sulfate, approximately 10% of all expressed genes have been upregulated. 3% have been downregulated. With respect to λ-carrageenan and fucoidan, smaller fractions of the expressed genes have been upregulated (7 and 5%, respectively). Larger portions, 18% and 17% have been expressed at a lower degree. Generally, large portions of genes expressed have been linked to the respective substrate. For instance, 611 of 1500 expressed genes in case of chondroitin sulfate were exclusively expressed regarding this substrate. The focus of the gene expression analyses was set on potentially expressed sulfatases and FGEs.

Out of six predicted FGEs in *R.baltica* SH1^T (Gene IDs: RB4229, RB5028, RB8026, RB11498, RB11811, RB11998), one, RB9091, was found to be active in the presence of all sulfated polysaccharides, but not in glucose grown cells (Table 5). The formation of a catalytic active formylglycine residue is crucial for sulfatase activity. The observation of only one FGE being active in case of sulfated polysaccharides raises the question how sulfatases

expressed under reference conditions are maturated or whether they are active at all. A recently described alternative model of sulfatase maturation was found by knocking out known maturation systems in *E. coli* (242). Analogous knock out experiments would allow drawing conclusions regarding alternative maturation systems in *R. baltica* SH1^T. Since genetic tools for planctomycetes are becoming more and more available (272), respective experiments should be possible in the near future.

Characteristic sulfatase expression profiles were yielded relating to all substrates. In case of glucose, eight sulfatase genes were expressed, four arylsulfatases (RB4815, RB7875, RB3849, RB9091, RB9549) and four N-acetylgalactosamine-6 sulfate sulfatases (RB200, RB3403, RB198, RB9091). In previous transcriptome studies conducted by Wecker and colleagues, focusing on the life cycle of *R. baltica* SH1^T and potential stress responses, glucose also was the substrate of choice (126,252). Comparing sulfatase expression data from those studies with this study, revealed a rather small intersect of two commonly expressed sulfatases, RB3403 and RB4815. RB3403 was observed by Wecker and co-workers to be repressed 300 min. after heat shock induction. It was concluded, that RB3403 is maybe involved in morphological remodeling in response of heat stress. Possibly it is involved in restructuring or adapting the holdfast substance *R. baltica* SH1^T is known for. RB4815 was hypothesized to be involved in attaching to solid surfaces, thus being part of the machinery enabling a sessile lifestyle. Though six sulfatases were expressed in case of fucoidan, respective data are not considered since hardly any growth was seen for this substrate.

The sulfatase expression profile from λ-carrageenan was observed to be comparable similar to that from glucose with few exceptions. Two sulfatases active in case of glucose (RB198, RB 9549), were inactive in λ-carrageenan, instead two sulfatases were expressed, of which one (RB4787) was exclusively expressed in λ-carrageenan grown cells.

Referring to chondroitin sulfate, 14 sulfatases were shown to be active, two N-acetylgalactosamine-6 sulfate sulfatases (RB406, RB9091) with one (RB9091) being upregulated and 12 arylsulfatases (RB4815, RB1477, RB5146, RB7875, RB13148, RB2357, RB348, RB3849, RB9091, RB9755, RB5355, RB3177, RB5294) (Table 5). RB9091 was only active in case of chondroitin sulfate and λ-carrageenan and is so far functionally unknown from previous studies. Eight sulfatases have been exclusively expressed in chondroitin sulfate grown cells considering all tested substrates. Out of the mentioned 14 sulfatases, 5 were previously identified to be active under specific stress conditions (126). RB406, RB5146 and RB13148 were repressed in case of heat or cold shock conditions. RB1477 was observed to

be expressed relating to heat and salt stress. RB4815, already mentioned to be linked to attaching to solid surface, was also found to be active.

The interplay between sulfatases from *Rhodopirellula* and sulfated polysaccharides of marine origin

Any sulfatase gene whose expression was observed during growth studies of *R. baltica* SH1^T were conserved in at least one other species of this genus. This, however, is true for any sulfatase encoding gene is this strain, as it is not providing any exclusive gene (Figure 16A). All of the expressed sulfatase genes contain a single sulfatase domain, except RB9549, which consists of two fully developed sulfatase domains. Assuming an involvement in polysaccharide degradation, it is hard to deduce whether sulfate ester cleavage occurs inside or outside of the cells based on the prediction of signal peptides and transmembrane helices. Most sugar transport systems, like the PTS (phosphotransferase) system, are specialized for the translocation of monomers (273,274), which suggests that sulfate ester cleavage might occur outside or inside dependent on whether sulfate esters are cleaved at the di- or monosaccharide stage.

Physiological aspects of sulfatases besides polysaccharide degradation

Independent from the substrate, *R. baltica* SH1^T constantly expresses a set of three sulfatases (RB4815, RB7875, RB3849). Their constitutive expression moves the focus from sulfatases being solely involved in utilizing sulfated polysaccharides to further functions. Recently, Wecker and colleagues (2009, 2010) (126,252) deduced a couple of additional functions of sulfatases in *R. baltica* SH1^T based on transcriptional studies relating to changing environmental conditions and life cycle analysis. Some of these sulfatases were shown to be also active under the conditions investigated during this project.

Another possible metabolic role of sulfatases is the production and secretion of extracellular polymeric substances (EPS) including exopolysaccharides (EP). EPs occur in two different ways: Capsules, which are tightly attached to cell walls, and slime polysaccharides that are loosely attached. *R. baltica* SH1^T is known to produce a holdfast substance that enables attachment to surfaces and to form cell aggregates (235). The composition of the holdfast substance is so far unknown. Nevertheless, sulfated polysaccharides are common components of EPS and EP from marine bacteria (275). Sulfatases can play an important role in

configuring and reconfiguring eventually present sulfated polysaccharides in the holdfast substance of *R. baltica* SH1^{T}.

5.5. Conclusion and Outlook

The exceptionally high copy number of sulfatases within the nine planctomycetal genomes is an outstanding feature of these organisms. Such high numbers are normally only found for e.g. transporters or regulators. The bioinformatic analysis of 1120 sulfatases revealed 240 discriminable lineages of exclusively Cys-type group I sulfatases, grouping into 19 major phylogenetic clusters. No Ser-type sulfatase has been found in the data. Only for five of these clusters, well-described orthologues in other organisms are currently known. A core set of 60 sulfatases occurring in all nine investigated organisms has been identified, as yet are of unknown function, but represent prime targets for future experimental analysis. We suspect sulfatases having cellular functions within these 60 selected ones. The distribution of sulfatases in examined strains reflects the phylogentic distance between those strains. We interpret the huge diversity of sulfatases as a response to the diversity of sulfated compounds in nature and especially in the marine environment. For *R. baltica* SH1^{T}, distinct sulfatase expression profiles in cells grown on different sulfated polysaccharides proved a functional link between sulfated polysaccharides and planctomycetal sulfatases. In line with previous studies and the constitutive expression of a subset of sulfatases points towards a central role in cellular functions beyond polysaccharide degradation.

Acknowledgments

We would like to express our gratitude to Andreas Ellrott and Emina Karamehmedovic for help during microarray processing and laboratory assistance. Many thanks to Gurvan Michel for detailed information about sulfated polysaccharides in marine environments. Thanks a lot to Florian Battke for straightforward help relating to MayDay. This project was funded by the Max Planck Society, which we gratefully acknowledge.

6. Summary and discussion

RNA based research can address a wide range of research questions demonstrating the miscellaneous functions of RNA molecules. The work accomplished during this thesis focused on the role of the 16S rDNA as a marker gene for biodiversity and mRNA transcripts as information donors for cellular functions. Although both approaches differ in terms of methodology and application, they share a common ancestor within this thesis: the MIMAS project. In this large-scale multi-'omic' project, the succession of the bacterioplankton population at Helgoland Roads in the North Sea was successfully characterized. Different 'omic' approaches, as well as classical 16S rDNA biodiversity studies, provided the basis to unravel the taxonomic and functional potential of the marine community when adapting to a spring bloom situation. Besides the cultivation independent approaches, the impact of pure culture studies brought us one step closer to unraveling the mysteries behind high amounts of sulfatase encoding genes in marine genomes, which should not be underestimated.

Taken together, in this thesis RNA based research demonstrated its attractiveness and power with respect to functional and taxonomic characterization. The results of the MIMAS project provide a wealth of new hypothesis and ideas for follow-up studies. The established research network and infrastructure provides a comprehensive basis for different kinds of studies to solve the big puzzle in understanding the unseen microbial world.

6.1. Evaluation of 16S rDNA primer and primer pairs

To start with, PCR based 16S rDNA diversity studies using next generation sequencing (NGS) technologies appeared an attractive and useful step to gain a quick overview of the phylogenetic composition at Helgoland Roads. However, the evaluation of primer and primer pairs was necessary to assure a minimum biased picture of the diversity on site. This issue has been addressed by the *in silico* evaluation of 175 primers and 512 primer pairs with respect to overall coverage and phylum spectrum for *Archaea* and *Bacteria*. For this purpose, the primer sequences were compared to the 16S/18S rDNA sequences in the SILVA non-redundant reference database (SILVA SSURef NR) release 108. To ensure a broad coverage of commonly used universal 16S rDNA primers, sequences were either obtained from a literature survey or kindly requested from the SILVA user community. Thereby, it became obvious that several differently named primers exhibit the same sequences, indicating the need for a

standardized nomenclature. In order to address this issue, all analyzed primers were renamed according to the nomenclature originally introduced by Alm et al. (139). By sending all essential primer information to probeBase (276), a central platform has been created to provide the scientific community with an overview of commonly available primers. Thereby, scientists might leave their routines behind and be more willing to use different primer pairs, which could provide more reliable results. A central domain also persuades with the advantages of easy maintenance and actuality. The existing options for submission of probes, even prior to publication, could be easily extended for primer submission. This would ensure an up-to-date database and remain attractive for the future.

The screening of the single primers provided a first overview about the sensitivity and specificity of the commonly used primers. It soon became clear, that one third failed to pass our threshold of 50% overall coverage, indicating the urgency of re-evaluation. However, the remaining two thirds convinced with relatively good results. Using a 75% overall coverage criterion, 86 single primers qualified for primer pair evaluation in the end. In-depth analysis included overall coverage, phylum spectrum, mismatch position and amplicon length. Based on the achieved results, the scientific community was provided with a set of 10 recommended archaeal and bacterial primer pairs suitable for different NGS technologies as well as the classical cloning and sequencing approach. The whole study is intended to serve as a guideline for finding the most appropriate primer pair for diversity studies in any habitat using any sequencing method. Unfortunately, the evaluation also revealed the urgency of designing new *Archaea* specific primers. Not only did the majority of *Archaea* specific primers fail to detect *Nanoarchaeota* but also gained disappointing results in terms of long amplicons. However, taking into account that intensive sequencing efforts revealed several novel archaeal taxa in the last couple of years, these results are not surprising. The majority of primers were designed prior to knowledge of the present diversity. This indicates not only the urgency in terms of re-evaluation but also the on-going need to design new primers. With the permanent exploration of the microbial world, it is not surprising that soon or later any primer has to deal with deficiencies.

The effort in setting up the informatic infrastructure for the primer evaluation also provided the basis for the development of the SILVA TestPrime tool as part of the SILVA website, which allows performing an online *in silico* PCR on the SILVA database with a primer pair of interest. Coverage is given for each taxonomic group, making it easy to quickly identify strengths and weaknesses of a particular primer combination. Moreover, selection between different SILVA datasets and mismatch conditions allows a flexible evaluation. This tool

clearly provides a valuable application, and it is desirable that it appeals to the scientific community. Pre-evaluation of newly designed primer pairs could ensure a more reliable picture of the microbial community and prevent failure detection of important key players. Last but not least, previously recommended combinations can be easily re-evaluated as soon as a new database release becomes available. One has to keep in mind that with the ongoing sequence accumulation in the databases, the re-evaluation of primers will always have a crucial role in the future.

Furthermore, the expectations are high that intensive metagenomic studies will provide more and more PCR-free high quality reads which will help to determine, and finally reduce, the postulated primer bias in the current rDNA databases. Although, the GOS dataset can be seen as a fruitful basis, it still contains too few numbers of sequences. Taking into account that the SILVA project team already made a first step towards high quality databases by creating the SILVA SSU Ref database, it is desirable that they might create a new database consisting only of almost full length sequences derived from PCR-free methods in the future. Unfortunately the majority of metagenomes are currently sequenced using short length NGS technologies such as Illumina. However, the rising SMRT technology from PacBio clearly has the potential to address this backlog with respect to full length sequencing. Currently, the high error rates prevent scientists from using this platform. However, with ongoing technological improvements the hopes are high that NGS based full length 16S rDNA sequencing coupled with good accuracy will become feasible in the near future. Until then, the primer bias issue is expected to remain in the databases for a while.

6.2. NGS based 16S rDNA analysis – proof of concept

The previous primer evaluation was intended to serve as a guideline for choosing the best available primer pair. As a proof of concept, 454 pyrosequencing based on 16S pyrotag analysis has been applied within the MIMAS project. The intention was that it should serve as a screening tool to identify the bacterioplankton succession in response to a diatom bloom in the North Sea. The outcome of the PCR amplified 16S rDNA ‘pyrotags’ analysis clearly demonstrated its suitability by identifying the dominant members of the community. Simultaneously, biodiversity analysis based on metagenomics further confirmed the results.

Although the MIMAS project initially served as a guinea pig project, it was also among the first to benefit from the evaluation. 16S rDNA pyrotags clearly complemented the quantitative catalyzed reporter deposition fluorescent-in situ-hybridization (CARD-FISH) analysis by providing high resolution data down to the genus level allowing the identification of key

players like *Formosa* and *Candidatus Pelagibacter ubique* without prior knowledge. Moreover, unexpected taxa such as *Reinekea* have first been detected by 16S rDNA pyrotags and induced the design of a new probe for follow up quantitative FISH analysis.

The importance of careful evaluation has been demonstrated by the application of a suboptimal primer pair on the same samples. Although the three dominant taxa, *Flavobacteria*, *Alphaproteobacteria* and *Gammaproteobacteria*, could be detected, discrepancies arose on the genus level. In-depth analysis revealed that the second primer pair has to deal with a strong mismatch issue within the flavobacterial taxa. In particular, the dominance of *Formosa* could only be detected to a minor extent. Although it is widely accepted that a standard PCR can tolerate one to two mismatches, the results confirm that it can easily result in a biased picture of the diversity. It is important not to underestimate this issue, and to always carefully check the primer target positions. A mismatch towards the 5'end appears to be unproblematic as demonstrated in the experimental evaluation; however, it is difficult to draw the line at what position a mismatch may interfere with the amplification. Therefore, we are more in favour of those primer and primer pairs which exhibit a high coverage and wide phylum spectrum coupled with low amounts of mismatches. Notwithstanding all limitations, NGS based 16S rDNA analysis clearly showed its suitability as a high resolution screening tool, and hence, can play a central role in future ecosystem monitoring.

6.3. Functional analysis of the bacterial community at Helgoland Roads in the North Sea

The MIMAS project aimed at investigating the bacterioplankton response to an algae bloom in early spring at the long-term ecological research site (LTER) Helgoland Roads. The taxonomic assessment, as briefly described in the previous section, provided a fundamental basis for-follow up studies. Next, a joint attempt of all MIMAS partners and different 'omic' approaches were applied to access the genetic and functional potential of the free-living fraction. In the beginning, significant efforts in designing and improving the experimental pipeline had to be made. This included a weekly sampling procedure from winter till autumn, which resulted in an extraordinary bio-archive containing multiple samples per sampling day. In addition, a bioinformatic pipeline for identification and annotation had to be created. In the end, those efforts led to a robust infrastructure for the performance of large-scale projects, and resulted in comprehensive data sets.

Initial 16S rDNA screening coupled with CARD-FISH helped in picking the most promising sampling dates for the different 'omic' approaches. In particular with the occurrence of the spring algae bloom a very dynamic bacterial succession could be observed. Briefly, members of the alphaproteobacterial SAR11 clade dominated the community in the pre-bloom phase. With the occurrence of the spring algae bloom, a change in the community structure resulted in an increase of *Flavobacteria* with high abundances of the *Formosa* and *Polaribacter* species. *Gammaproteobacteria* members showed a constant increase in abundance and provided surprising results by the sudden appearance of *Reinekea*. Interestingly, setting the initial winter abundance as reference, the termination of the spring phytoplankton bloom boosted microbial cell densities to a seven fold higher level by the end of April, followed by a subsequent decline to a threefold initial level in early May. During this period, a profound dynamic composition change in the bacterial community could be observed. Due to this interesting taxonomic pattern samples from the pre-, inter- and early post-algae phase qualified for in depth functional characterization by the application of a full meta-'omic' approach.

Metagenomes not only gave first insights into the genetic potential but also provided the backbone for the identification of metaproteome and metatranscriptome results. On average 65% of the metatranscriptome reads and metaproteome fragments could be mapped to ORFs of the metagenome dataset, demonstrating its essential part within a multi 'omic' study. The coupling of gene content with functional analysis further allowed investigations with respect to gene shift and expression. For example, taxonomic distinct expression of transporters, glycosyl hydrolases (GH) and sulfatases have been detected and provided first evidence of pronounced nutrient strategies of the dominant taxa. Noticeable is the high amount of transcripts encoding for sulfatases and GH families (in particular GH16 and GH13) indicating that *Flavobacteria* are able to degrade complex polymers from algae. Increasing transcripts encoding for TonB dependent transporter (TBDT) components are suggested to be involved in uptake of complex polysaccharides. Further evidence provided genomic studies of marine metagenomes, which revealed that the TBDT genes are often co-localized with carbohydrate degrading enzymes. Based on this distinct expression profile and the increasing abundance in the early spring bloom we believe that *Flavobacteria* are the first to benefit from the changing nutrient conditions. The degradation of the complex polysaccharides resulted in an increasing availability of sugar oligomers and monomers. The latter are of particular importance for opportunistic bacteria like *Roseobacter*, which was also reflected by the membrane transporter profile. With a high expression of monomer transporters such as ABC transporter,

coupled with solute binding proteins (SBP), *Rosebacter* and *Alphaproteobacteria* dominated the uptake of monomers. This also complies with the ecological strategies of SAR11 clade members, who are believed to thrive on low nutrient conditions by the high expression of low molecular weight transporters coupled with proteorhodopsin. The latter has also been identified within the metatranscriptome, suggesting that it provides a fitness advantage for SAR11 clade members to quickly regain their dominance as soon as the algae bloom dominating taxa are starting to vanish. In summary, the results gained from the 'omic' approach provided a first insight in how the bacterioplankton members can evade extinction despite their seemingly homogeneous habitat.

The core of abundant transcripts has been simultaneously detected by the metaproteome and metatranscriptome confirming its expression with high confidence. However, small discrepancies arose with respect to taxonomic resolution and expression level. For example, *Roseobacter* featured a greater expression of membrane transporters in the metatranscriptome than in the metaproteome. Moreover, only a few mRNA transcripts encoding for phosphate and phosphonate transport systems could be detected. It is still unclear why those differences in the expression profiles occurred, but it has been suggested that it might be a result of posttranscriptional regulation, mRNA degradation or fast transcriptional adaptation to an external stimulus.

The discrepancy between metatranscriptomics and metaproteomics also highlighted the need for a combined approach to fully unravel the complexity of the functional patterns within a microbial community. However, due to financial constraints and a lack of expertise, very few research groups employed more than one 'omic' approach in the past. Obviously, no individual group has the resources or capacity to perform a comprehensive characterization of a complex environmental sample. Therefore, it is of substantial need to intensively collaborate in large-scale 'omic' projects to combine the expertise of several research groups.

Metatranscriptomics further excels by providing complementary information. Thereby, surprising results with respect to the photosynthetic activity have been revealed. Interestingly, the striking peak (14.04.2009) of photosystem encoding transcripts was assigned to *Cyanobacteria*. To our surprise, only a very low amount of cyanobacterial 16S rDNA sequences could be detected. Rather, a comparatively high number of *Chloroplast* 16S rDNA sequences dominated the sample. Unfortunately, in this particular case, clear taxonomic assignment of the mRNA transcripts was impossible. The NCBI taxonomy lacks distinction between *Chloroplast* and *Cyanobacteria* as available for ribosomal RNA in SILVA. Although it is suggested that this peak most probably originated from algae, the striking increase was

unexpected and calls for further in-depth transcriptomics of the whole phytoplankton community.

Furthermore, the simultaneous detection of cDNA derived from mRNA and rRNA provided first insights into the active adaptation of microbes to changing nutrient conditions. Interestingly, *Roseobacter* mRNA transcripts appeared to degrade more rapidly while rRNA transcripts remained highly abundant. These results provided the first evidence that members of this taxa adapt rapidly to changing nutrient availabilities. Taking into account that metabolic active bacteria contain a higher amount of rRNA transcripts, detection of 16S rRNA transcripts can serve as a marker for the fitness status of microbial communities in the future. Moreover, the number of rRNA encoding genes per genome appears to correlate with the time how fast microbes are able to react to altering nutrient conditions. For example, members of the taxa *Reinekea* have up to four rRNA operons, which might explain its sudden appearance.

In summary, the MIMAS project successfully demonstrated that it is possible to connect the expertise from different research groups, and to combine microbial diversity studies and functional 'omic' data into a very detailed *in situ* analysis. The study showed that the bacterial response to an algae bloom at Helgoland Roads is much more dynamic than previously anticipated. Distinct populations e.g. *Flavobacteria* appeared to play an active role in in algae decomposition and thus carbon turnover. The results further suggested that the bacterioplankton dynamics resulted from the successive nutrient availability of different algae primary components (bottom-up control). Thereby, ecological niches were provided and allowed specialized taxonomic groups to grow. The results can help to unravel the 'truth' about how bacterioplankton members prevent extinction by direct competition.

In combination with the contextual data from the LTER, the gained results have the potential to serve as a basis for building general theories and principles. They could result in predictive models for bacterioplankton bloom dynamics and thus provide deeper insights about ecological niche defining factors. Moreover, the simultaneous detection of cDNA derived from mRNA and rRNA could serve as a fitness marker to determine the health status of a bacterial community.

6.4. Impact of pure culture experiments

Beside the gained biological conclusions, an extensive laboratory effort has been made to set up the experimental pipelines. In particular, RNA extraction, RNA clean-up and cDNA synthesis had to be further optimized, leading to high and mostly unbiased RNA

concentrations; technically, even small regulatory (sRNA) transcripts remained in the sample. sRNA are crucial regulators of the prokaryotic gene expression and the molecules range from 50 to 250 nt in length. The commercially available clean-up procedures using columns usually cannot bind fragments smaller than 100 bp. However, with a small modification in the RNA clean-up protocol, the loss could be prevented.

To avoid an unnecessarily high use of valuable filters from the bio-achieve, *R. baltica* SH1^T was used as a model organism for the establishment and fine tuning of the molecular techniques. Although this bacterium was isolated in the Baltic Sea, closely related species covered the eastern North Sea (254). It soon became clear that a pure culture from the marine habitat is essential for setting-up the experimental pipeline. For example, breaking up the cell wall of marine microbes to release the nucleic acids appeared to be very challenging. Finally, combined mechanical and chemical cell lysis achieved the best results. Although it appeared to be a simple step in the end, we would have most likely failed without prior experience from the work with marine model organism.

Undoubtedly, model organisms are not only a useful tool for method development but also allow studying and answering questions that are currently impossible to answer on the environmental level. In addition, cultivation-independent meta-'omic' projects such as MIMAS often provide a wealth of ideas for follow up experiments. The functional analysis of the bacterial community particularly revealed a peak of CAZyme expression being accompanied by sulfatases; the latter especially attracted our attention. In the marine environment, sulfated polysaccharides from algae are believed to be cleaved by sulfatases, and hence present potential substrates for a variety of microbes. However, little is known about substrate specificity of sulfatases and its distinct role in metabolic and cellular processes. This indicated the need for further in-depth analysis on pure cultures and *R. baltica* SH1^T excels as a suitable model organism due to its high amount of potential sulfatase encoding genes.

Growth experiments on different sulfated substrates revealed the first evidence of potential substrate utilization. Interestingly, chondroitin sulfate resulted in a higher growth rate than the reference sample, indicating that this substrate is a better match to the natural environmental than glucose. In the future, the improvement of standard media for culturing of *R. baltica* SH1^T might be necessary to get one step closer towards natural conditions.

Follow-up gene expression analysis provided a first functional link between sulfated polysaccharides and planctomycetal sulfatases and allowed speculations of potential substrate specificity. Bioinformatic assessment further interpreted the diversity of sulfatases as a

response to the wide variety of sulfated compounds in the marine habitat. However, a constant expression independent from the substrate availability indicated diverse functions far beyond polysaccharide degradation. It has been suggested, that they might play an important role in the remodelling of the distinct morphological features of *R. baltica* $SH1^T$ such as the characteristic holdfast substance (126,252).

This study can be seen as a basis for future follow-up studies and generating new hypothesis. It is recommended to perform knock-out experiments with *R. baltica* $SH1^T$ (272) to draw further conclusions with respect to substrate utilization, sulfatase maturation systems and cellular functionality. To gain a functional link between the expressed sulfatases of *Flavobacteria*, as seen in the MIMAS metatranscriptome, and potential substrates, the culturing of key players such as *Formosa* are requested. Further investigations concerning sulfatase function and substrate specify could clearly benefit from the present bioinformatic and laboratory infrastructure as well as the data and biomass archive.

7. Conclusion and outlook

The applied multi ‘omic’ approach has shown to be a powerful approach for the characterization of microbial communities and provides the scientific community with a wide variety of follow-up experiments and ground-breaking new hypothesis. The gained results within this thesis, and in particular the MIMAS project, provide a promising basis for future applications and research aims implemented in a complex network.

7.1. Recycling of the accumulated data and bio-archive

One of the first steps of the MIMAS project included the setup of a weekly sampling procedure from early spring until late autumn in the year 2009. Obviously not all sampling dates could be addressed within the multi ‘omic’ study. This meant that a large bio-archive accumulated and is ready to be used in follow-up studies. For example, as soon as new sequencing technologies become commercially available, the remaining samples can be used for re-analysis or extended functional profiling. The outcome could be of substantial use to validate and adjust new methods as well as to identify potential biases and bugs. Furthermore, other completely different methods can be applied to gain additional information to further characterize the microbial community at Helgoland Road, for example, diversity studies based on the 23S rDNA gene or GeneFISH (277) to link gene presence with cell identity. In addition, functional and taxonomic characterization of the particle-attached bacterial community might provide further evidence about algae colonization or lysis.

An intensive laboratory effort has also been invested in setting up the experimental metatranscriptomic pipeline. Besides generating high yields of RNA samples, it was of particular interest to keep sRNAs. Fine-tuning of RNA clean-up and cDNA synthesis technically prevented the loss of regulatory molecules. Due to the lack of time and missing expertise within the MIMAS consortium, in-depth analysis of sRNA could not be performed. However, in theory the dataset contains sRNA molecules, and is available for future analysis.

7.2. Generating guidelines and sticking to standards

Considerable progress has been made in providing the scientific community with a guideline for the choice of optimal primers. It would be desirable that the scientific communities accept those recommendations for future applications to assure more comparable results. Moreover it

would be wishful thinking to generate some sort of standards such as an agreement on the minimum amplicon length or coverage of hyper variable (HV) regions. Recently, the Genomic Standards Consortium (GSC) (278) began to drive a community-based standardization to maximize the outcome and comparability of sequencing data. For example, as part of the Micro B3 project (http://www.microb3.eu) all partners are expected to confirm to the minimum information standards published in the course of the GSC for describing samples as part of the Ocean Sampling Day in 2014. An analogue approach would be desirable for any 16S rDNA pyrotag screening. The Earth Microbiome Project already made a start by providing a standard protocol for 16S rDNA amplification using Illumina. However, with the rapid development of sequencing technologies, who can say what the 'golden procedure' will be in the near future? Currently, the best alternative is to standardize the documentation of the process as recommended by the 'Minimum Information about a MARKer gene Sequence' (MIMARKS) checklist (114), which excels with its universality with respect to target sequence and technology.

7.3. MIMAS on a global level and an open-access policy

MIMAS successfully characterized a bacterioplankton community at the well-known long term ecological research site (LTER) Helgoland Roads. The multi 'omic' approach demonstrated its power and it can only be recommended to apply this type of project on a world-wide scale. For example, samples taken within the ocean sampling day on midsummer (June 21st) in the year 2014 could be used to address different research questions. The currently planned 16S rDNA screening approach could be extended towards a full 'omic' study. Imagine the impact on marine genomics by combining microbial diversity studies and different 'omics' to access the genetic and functional potential of marine ecosystems around the world. This unique dataset would be fundamental for ecological forecasting and could result in knowledge that serves science and society for decades. However, long-term data storage and an open-access policy are required in the same breath. Modern research benefits from high circulation and the transfer of scientific knowledge, without doubt. In particular, open-access publication will help to avoid duplication and in the meantime accelerate innovations and efficiency. Recently, with the release of open-access, online magazines such as PLoS ONE the scientific society made one step forward towards making the world's scientific publications a freely available resource. However, economic barriers still hinder the transition to a free available scientific library. In most cases, the cost for payment is shifted

from the subscriber to the author, which is usually paid by the institute. Unfortunately, not every research group can afford the fees, indicating the need for a clear funding policy.

Besides an open-access policy, the efforts for long-term storage and maintenance of datasets should not be underestimated. In general, funding is short-lived, and it is very difficult to get post-project funding for the maintenance of data infrastructure and accessibility. It is without any doubt that long-term datasets are useless if they are only available for an exclusive core group and/or are disposed after a few years of storage. Improving data policy with respect to open and long-term accessibility is desirable to provide scientific society with a new basis for unraveling nature's ecosystem functions.

7.4. Ecosystem monitoring and the genetic treasure box

Most likely complex ecosystems alter over decades until a disturbance permeates natural systems. Disturbances such as species invasions, flood, and storms can influence the ecosystem and interferes with the ecological balance. However, time lags between cause and effect in ecological systems are the rule and it is unlikely for one person to sense slow changes occurring over decades. This indicates the importance of environmental studies at LTERs such as Helgoland Roads. These research sites do not just focus on creating new data but also on documenting and maintaining the large number of unique long-term datasets, which play a key role in long-term ecological research. The results gained within the MIMAS projects provide another important cornerstone. The accumulated knowledge furnishes a basis for future ecosystem monitoring, detecting daily, seasonal and annual changes in the ocean. This type of fundamental research might provide scientists and society with a conceptual framework for the assessment and forecasting of environmental situations. At present, we live in a time of unpredictable and dramatic global change, in which the ability of ecosystems to buffer natural or anthropogenic influences (for example, eutrophication) is becoming a very difficult task. Therefore, the identification of key players and components may help to understand and maintain ecosystem stability. In particular, bioremediation, as in the development of tools for restoring natural systems, is desirable. This includes, for example, the identification and degradation of xenobiotics as well as the development of useful treatments for ecosystem recovery. In particular, marine metagenomics are expected to be a fruitful resource for novel enzymes. However, not only bioremediation benefits from this huge data archive but also diverse industries have also discovered the marine gene pool to be worth investing in. The hopes are high that the widely unexplored microbes in the ocean are a treasure box full of novel biocatalysts for diverse biotechnological applications. For example

cellulases recently attracted attention. These enzymes play a central role with biofuel production involving lignocellulose plant material (279). At the moment, inefficient decomposition and enzymatic degradation hampers fuel production, indicating the need for novel cellulases. Metagenomes from the marine habitat may contain interesting new candidates; thus, this has the potential to open this bottleneck. Imagining marine enzymes to have a high impact on environmental and also political issues such as alternative energy resources demonstrates the importance of marine 'omic' applications, and explains the excitement in this research field.

7.5. Next generation networking

The MIMAS project developed an infrastructure for integrated projects and successfully demonstrated the advantage of splitting up a large-scale project into smaller subgroups addressing different aspects. This is easier said than done. The coordination and balancing of individual and independent research groups is challenging and requires a well guided and organized network. As a result the required efforts, the results were more than satisfying demonstrating that the outcome is more than the sum of its part.

In the past, it was difficult for a scientist to connect with other scientists unless he or she went to a conference or meeting. Unfortunately, due to time and financial constraints, not every scientist can fly around the world to attend conferences, no matter in which country they take place. Luckily, today's standards, such as email or even skype, have already made a big step towards establishing an easy communication system around the globe. However, maintaining and keeping the communication alive is the hardest part. It is without controversy that research did and will benefit from fruitful collaborations. Unfortunately experience has taught us that those collaborations could quickly become a disappointment if communication dies away.

At present, we live in a time where social media such as Twitter, Google+ and Facebook are becoming more and more a part of our daily life, whether we embrace or reject the notion. Briefly, those 21st century social networks provide a web based platform for users to connect and share interests, activities or real-life relationships. For example, Facebook advertises with the slogan 'Giving people the power to share and make the world more open and connected'. Interestingly, the core of the idea can be transferred to scientific collaboration, which gives scientists the power to connect with other research groups, share expertise and expand knowledge. Sharing, communicating and connecting always play a central role, no matter whether they involve a social component or are of purely scientific in nature. So why not

transfer the general concept of the social networks to the scientific world out there? Young and experienced generations of scientists would clearly profit from any kind of networking. Platforms such as Facebook and Twitter have created a fundamental basis for keeping communication alive by being freely available and using different kinds of media. For example, photos, videos, blogs entries and news updates provide fruitful bases that can have the power to keep research open-minded, and thereby serve science and society. The major advantage is that social networks are free, easy accessible and allow extravagating cultural and country frontiers. Notwithstanding the on-going discussion about the negative impacts, data protection and time-consuming factors, those platforms serve as a good way to make, exchange and maintain contacts. So why not use the basic concept behind social networks and transfer this concept towards a next generation science network? For example, participants of the Ocean Sampling Day could join a 'group of interest' and easily contact each other to share their experience, expertise and problems. Even if certain members are less active than others, the communication stays alive by the input of active community members and uploading of photos, videos and other related news. The use of different kind of media can be further recycled for public relations, and thus, provided to the society. This fruitful networking could also help to solve problems with respect to the bioinformatic and experimental set-up, which can lead automatically to a more standardized procedure. The latter assures more comparable results and brings scientific society one step closer towards an agreement upon standards in the near future. Moreover, the basic concept of networking stimulates interaction with other interested parties, which can lead to follow-up collaborations. Thereby new ideas for new or follow-projects can arise or the results from different projects can be connected. For example, the outcome from the MIMAS project can be linked to the outcome of the 2009-2010 Research Voyage of the Sorcerer II Expedition (http://www.sorcerer2expedition.org), which included samples from Helgoland Roads in June 2009. Moreover, interested parties can be found to expand the expertise of follow-up projects such as a new biotechnological part within MIMAS II. The hopes are high that flexible networking maintains and expands its positive impact on the scientific dialogue and society in the near future.

8. Acknowledgments

To begin with, I would like to thank Frank Oliver Glöckner for accepting me as a biologist to the Bioinformatic research group, for giving me the opportunity to work in such an interesting research field, for being an excellent supervisor with great leadership skills and – most important – for giving me the freedom I needed to fully develop my skills.

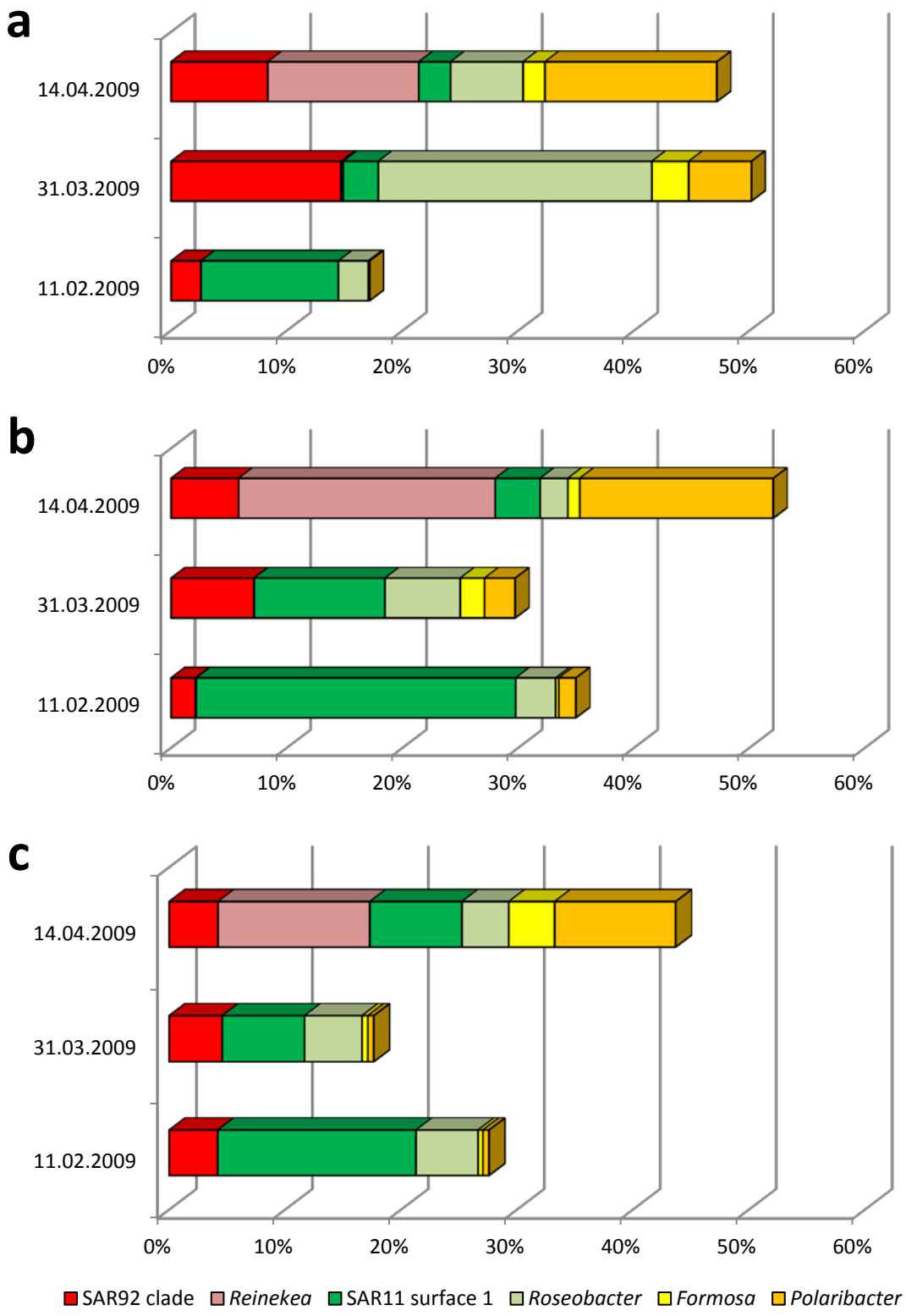
I would also like to thank Jack Gilbert and Matthias Ullrich for being in my thesis committee and for the helpful discussions in my thesis committee meetings.

Special thank goes to the Microbial Genomics and Bioinformatic Research Group (MGG) and the Molecol for their support and great working environment. In particular, I would like to thank Christine Klockow, Elmar Prüsse, Alexander Mann, Emil Ruff, Harry Potter, Jost Waldmann and Matthias Winkel. Not to forget my Marmic friends such as Amandine Nunes-Jorge, Petra Pjevac, Mar Fernandez Mendez, Great Reintjes and Juliane Wippler. Special thanks also to Emina Karamehmedovic and Carl-Eric Wegner.

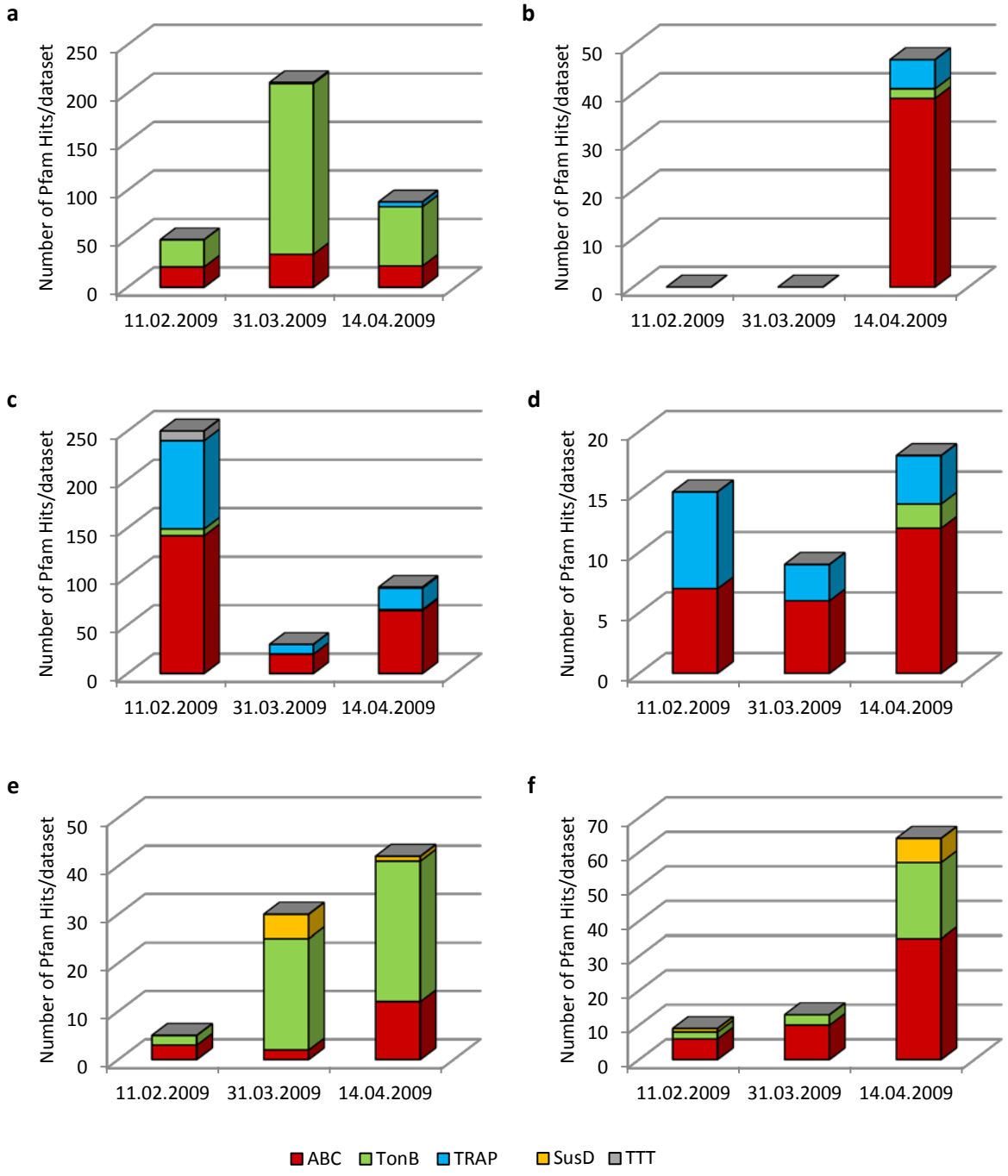
Last but not least, I would like to thank Matthias Kopf, my family, friends and PolyA-Alva Rosendahl for their continuous support throughout my whole life.

May the force be with you!

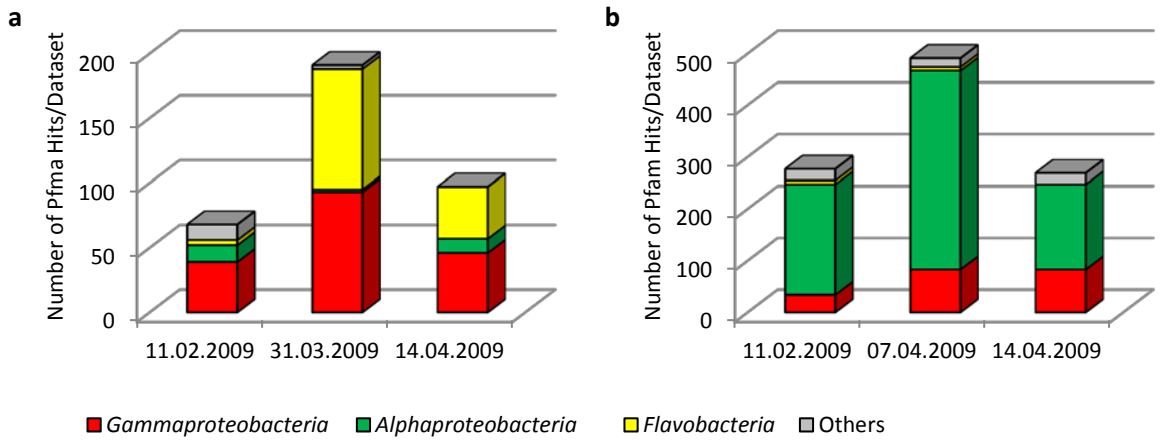
9. Supplementary Material



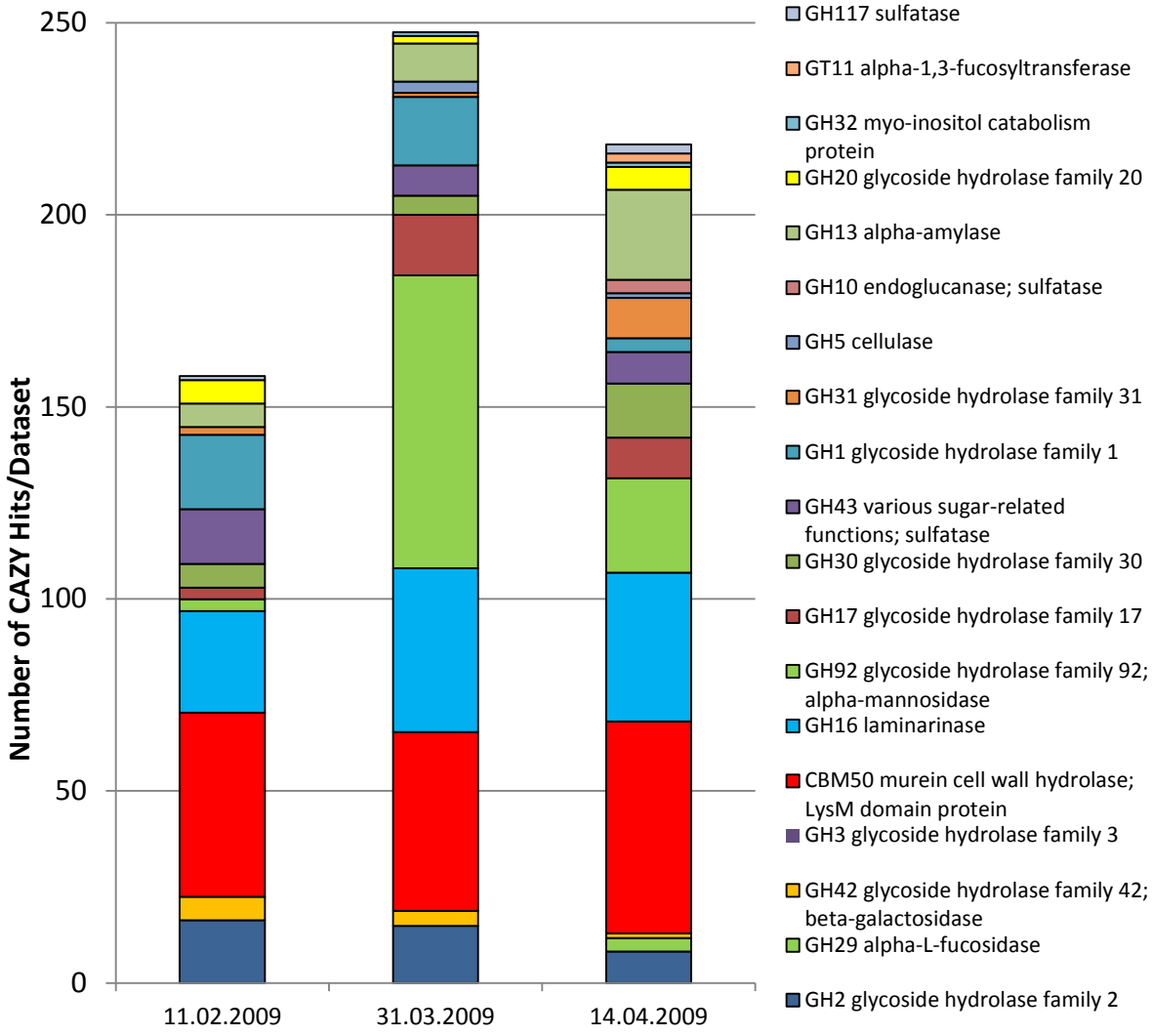
Supplementary Figure 1: Taxonomic profile of six dominant taxonomic groups. 16S rDNA reads were gained from a) directly sequenced cDNA (16S RNA), b) PCR amplified pyrotags (16S pyrotags) and from c) metagenome (16S metagenome).



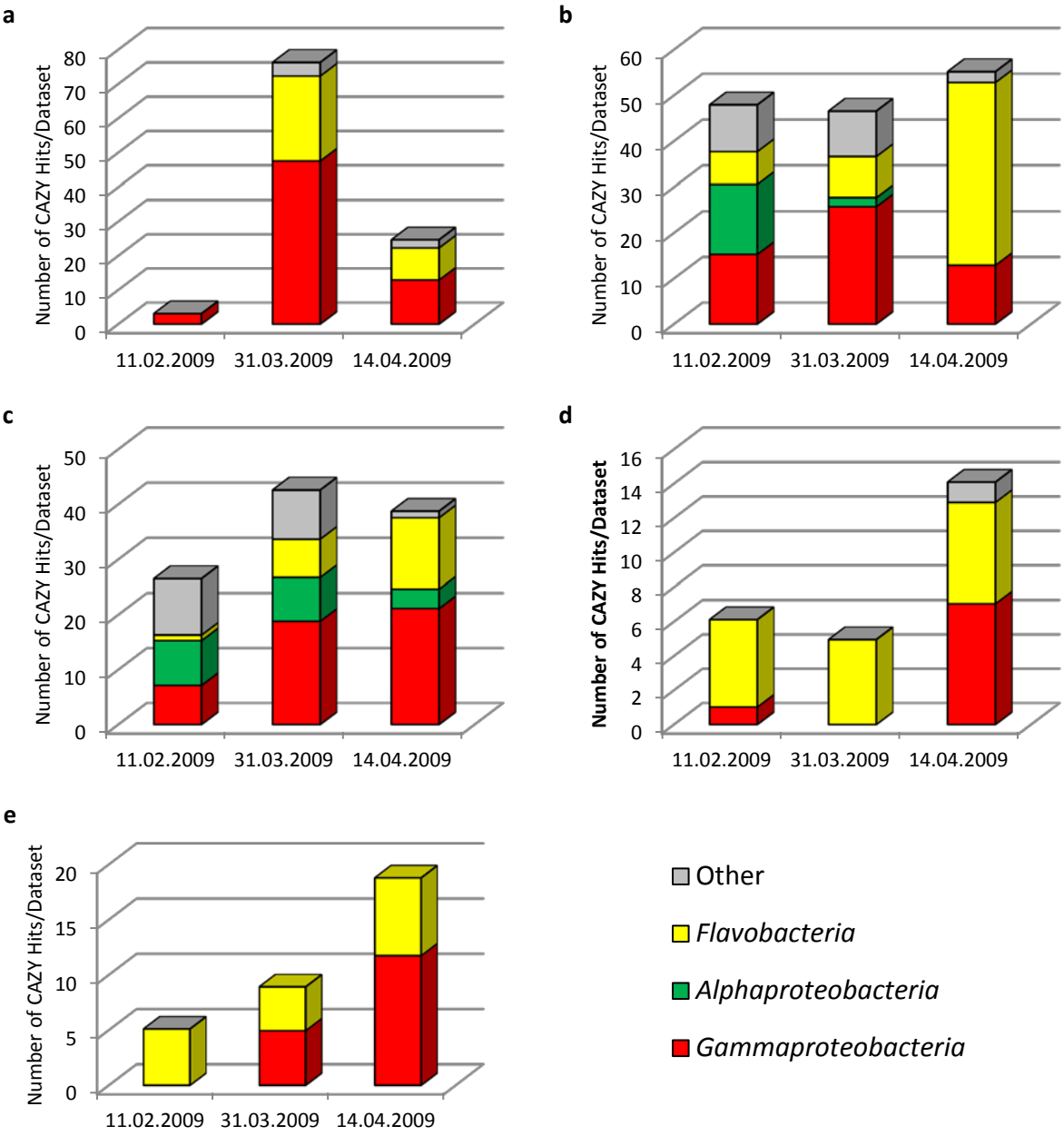
Supplementary Figure 2: Pfam annotations of transcripts encoding for membrane transporter for six abundant taxonomic groups a) SAR92 clade, b) *Reinekea*, c) SAR11 clade, d) *Roseobacter*, e) *Formosa* and f) *Polaribacter*. The abbreviations for the transporters are: TonB-dependent transport systems (TBDT), starch utilization system proteins (SusD), ATP binding cassette (ABC), tripartite ATP independent (TRAP) and tripartite tricarboxylate transporters (TTT).



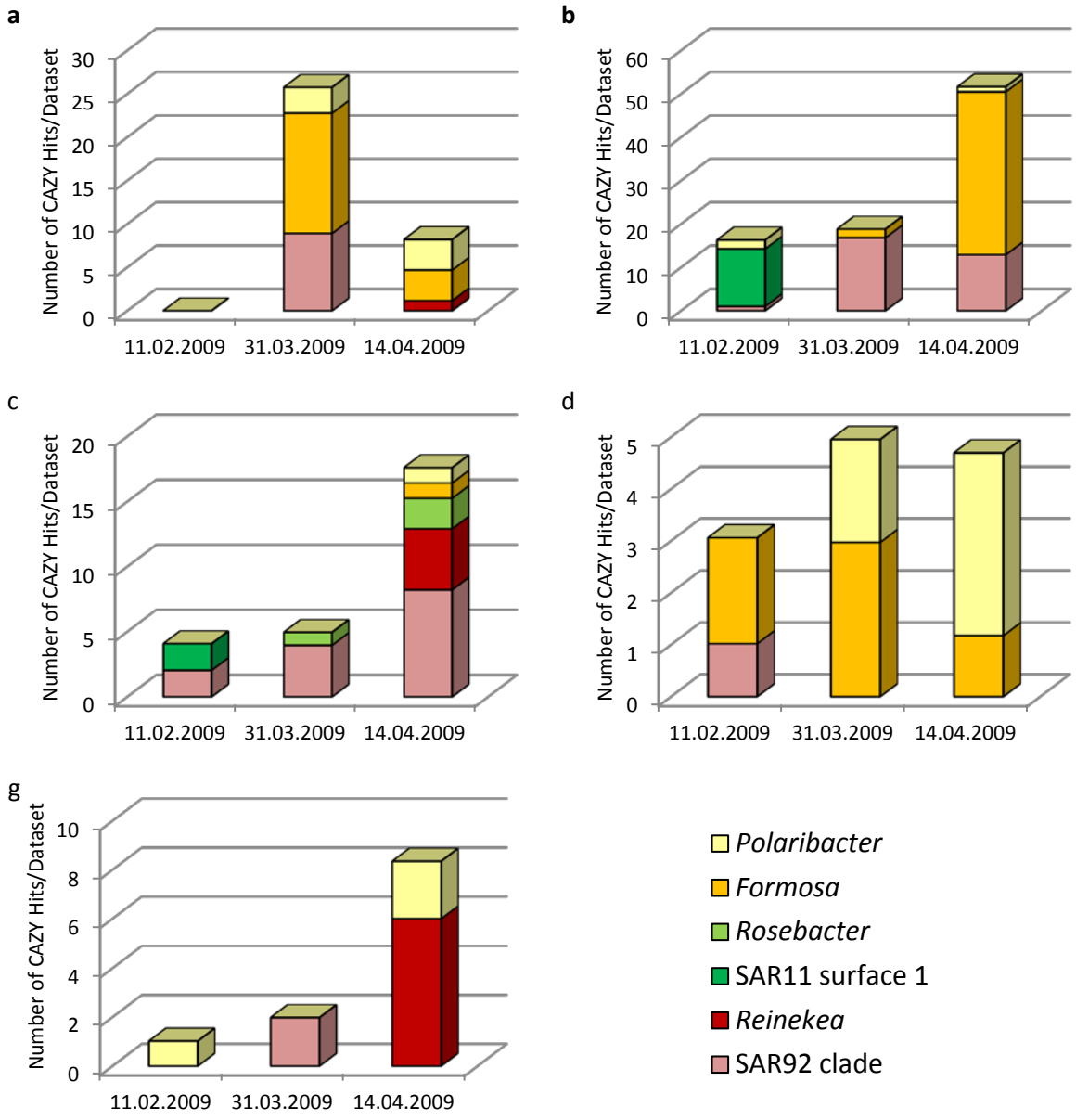
Supplementary Figure 3: Pfam annotations of transcripts encoding for a) cytoplasmic transmembrane components of the TonB complex (ExbB and ExdD) and b) bacterial extracellular solute-binding proteins (SBP)



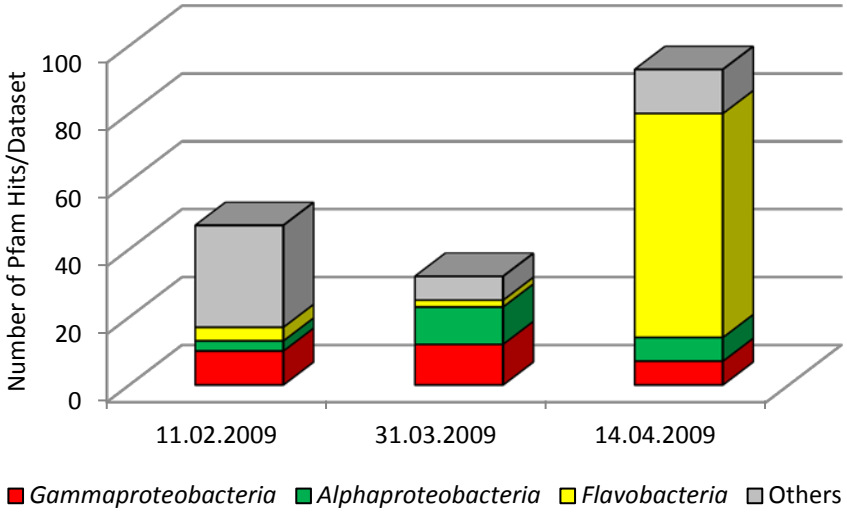
Supplementary Figure 4.: Number of transcripts encoding for prevalent CAZymes involved in external carbohydrate degradation.



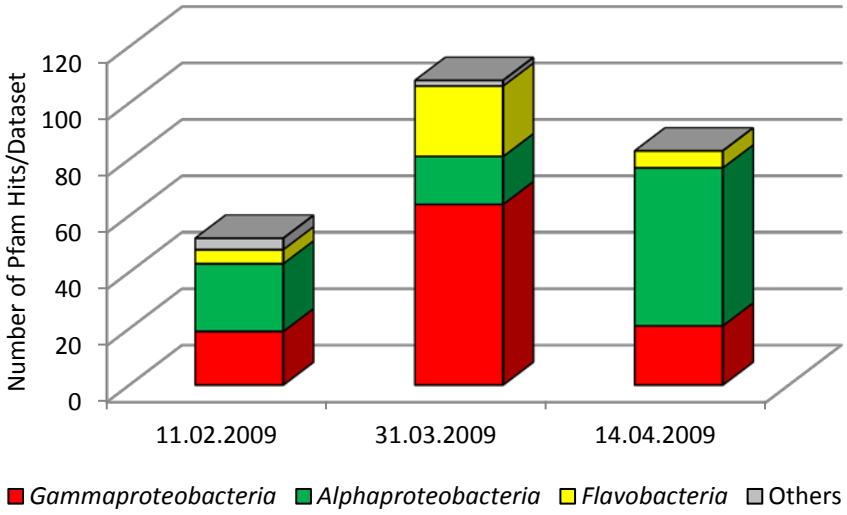
Supplementary Figure 5.: Number of transcripts encoding for selective CAZymes involved in external carbohydrate degradation: a) GH16, b) GH3, c) CBM50, d) GH30, g) GH13



Supplementary Figure 6.: Number of transcripts encoding for selective CAZymes involved in external carbohydrate degradation. Only CAZymes expressed by either SAR92, *Reinekea*, SAR11, *Roseobacter*, *Formosa* or *Polaribacter* were taken into account. CAZymes which were assigned to other taxonomic groups were not included: a) GH16, b) GH3, c) CBM50, d) GH30, g) GH13



Supplementary Figure 7: Pfam annotations of sulfatase encoding transcripts.



Supplementary Figure 8: Pfam annotations of Proteorhodopsin encoding transcripts.

Supplementary Table 1: Taonomic assignment of cDNA reads

taxonomic path	**11.02.2009**	**31.03.2009**	**14.04.2009**
Gammproteobacteria	21%	46%	31%
Alphaproteobacteria	40%	21%	17%
Flavobacteria	8%	15%	20%
Other	31%	18%	31%

10. References

1. Watson JD, C.F. (1953) A Structure for Deoxyribose Nucleic Acid. . *Nature*, **171**, 737–738.
2. Clancy S. (2008) Chemical Structure of RNA. *Nature Education*, **1**.
3. Moore SD, S.R. (2007) The tmRNA system for translational surveillance and ribosome rescue. *Annu Rev Biochem*, **76**, 101-124.
4. Slomovic S, P.V., Schuster G. (2008) Detection and characterization of polyadenylated RNA in Eukarya, Bacteria, Archaea, and organelles. *Meth Enzymol*, **447**, 501-520.
5. Sarkar N. (1997) Polyadenylation of mRNA in prokaryotes. *Annu Rev Biochem*, **66**, 173-179.
6. Gloeckner FO, K.M., Bauer M, Teeling H, Lombardot T, Ludwig W, Gade D, Beck A, Borzym K, Heitmann K, Rabus R, Schlesner H, Amann R, Reinhardt R. (2003) Complete genome sequence of the marine planctomycete Pirellula sp. strain 1. *Proc Natl Acad Sci U S A*, **100**, 8298-8303.
7. Fuerst JA. (1995) The planctomycetes: emerging models for microbial ecology, evolution and cell biology. *Microbiology*, **141**, 1493-1509.
8. Vogel J, S.C. (2005) How to find small non-coding RNAs in bacteria. . *J Biol Chem*, **386**, 1219-1238.
9. Storz G, H.D. (2007) A guide to small RNAs in microorganisms. . *Curr Opin Microbiol*, **10**, 93-95.
10. Vogel J, W.E. (2007) Target identification of small noncoding RNAs in bacteria. . *Curr Opin Microbiol*, **10**, 262-270.
11. Repoila F, D.F. (2009) Small regulatory non-coding RNAs in bacteria: physiology and mechanistic aspects. . *Biol Cell*, **101**, 117-131.
12. Griffin BE. (1971) Separation of 32P-labelled ribonucleic acid components. The use of polyethylenimine-cellulose (TLC) as a second dimension in separating oligoribonucleotides of ‘4.5 S’ and 5 S from E. coli. *FEBS Lett*, **15**, 165-168.
13. Ikemura T, D.J. (1973) Small ribonucleic acids of Escherichia coli. 1. Characterization by polyacrylamide-gel electrophoresis and fingerprint analysis. *J Biol Chem*, **248**, 5024-2032.
14. Silva IJ, S.M., Dressaire C, Domingues S, Viegas SC, Arraiano CM. (2011) Importance and key events of prokaryotic RNA decay: the ultimate fate of an RNA molecule. *Wiley Interdiscip Rev RNA*, **2**, 818-836.
15. Barrangou R, F.C., Deveau H, Richards M, Boyaval P, Moineau S, Romero DA, Horvath P. (2007) CRISPR Provides Acquired Resistance Against Viruses in Prokaryotes. . *Science*, **315**, 1709-1712.
16. Sorek R, K.V., Hugenholtz P. (2008) CRISPR — a widespread system that provides acquired resistance against phages in bacteria and archaea. . *Nat Rev Microbiol*, **6**, 181-186.
17. Withers M, W.L., Dos Reis M. (2006) Archaeology and evolution of transfer RNA genes in the Escherichia coli genome. . *RNA*, **12**, 933-942.
18. Horner-Devine MC, C.K., Bohannan BJM. (2004) An ecological perspective on bacterial biodiversity. *Proc R Soc Lond B*, **271**, 113-122.
19. Fierer N, B.M., Nulton J, Salamon P, Lozupone C, Jones R, Robeson M, Edwards RA, Felts B, Rayhawk S, Knight R, Rohwer F, Jackson RB. (2007) Metagenomic and

small-subunit rRNA analyses of the genetic diversity of bacteria, archaea, fungi, and viruses in soil. *Appl Environ Microbiol*, **73**, 7059-7066.
20. Roesch LF, F.R., Riva A, Casella G, Hadwin AK, Kent AD, Daroub SH, Camargo FA, Farmerie WG, Triplett EW. (2007) Pyrosequencing enumerates and contrasts soil microbial diversity. *ISME J*, **1**, 283-290.
21. Andersson AF, R.L., Bertilsson S. (2010) Pyrosequencing reveals contrasting seasonal dynamics of taxa within Baltic Sea bacterioplankton communities. *ISME J*, **4**, 171-181.
22. Sogin ML, M.H., Huber JA, Mark Welch D, Huse SM, Neal PR, Arrieta JM, Herndl GJ. . (2006) Microbial diversity in the deep sea and the underexplored 'rare biosphere'. . *Proc Natl Acad Sci U S A*, **103**, 12115-12120.
23. Dewhirst FE, C.T., Izard J, Paster BJ, Tanner AC, Yu WH, Lakshmanan A, Wade WG. (2010) The human oral microbiome. *J Bacteriol*, **192**, 5002-5017.
24. Hamady M, K.R. (2009) Microbial community profiling for human microbiome projects: Tools, techniques, and challenges. *Genome Res*, **19**, 1141-1152.
25. Fang M, K.R., Motavalli PP, Davis G. (2005) Bacterial diversity in rhizospheres of nontransgenic and transgenic corn. *Appl Environ Microbiol*, **71**, 4132-4136.
26. Amann RI, L.W., Schleifer KH. (1995) Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiol Rev*, **59**, 143-169.
27. Tringe SG, v.M.C., Kobayashi A, Salamov AA, Chen K, Chang HW, Podar M, Short JM, Mathur EJ, Detter JC, Bork P, Hugenholtz P, Rubin EM. (2005) Comparative metagenomics of microbial communities. *Science*, **308**, 554-557.
28. Huse SM, D.L., Huber JA, Mark Welch D, Relman DA, Sogin ML. (2008) Exploring microbial diversity and taxonomy using SSU rRNA hypervariable tag sequencing. *PLoS Genet*, **4**, e1000255.
29. Woese CR. (1987) Bacterial evolution *Microbiol Rev*, **51**, 221-271.
30. Clarridge JE 3rd. (2004) Impact of 16S rRNA Gene Sequence Analysis for Identification of Bacteria on Clinical Microbiology and Infectious Diseases. . *Clin Microbiol Rev*, **17**, 840-862.
31. Olsen GJ, L.D., Giovannoni SJ, Pace NR, Stahl DA. (1986) Microbial ecology and evolution: a ribosomal RNA approach. *Annu Rev Microbiol*, **40**, 337-365.
32. Nossa CW, O.W., Yang L, Aas JA, Paster BJ, Desantis TZ, Brodie EL, Malamud D, Poles MA, Pei Z. (2010) Design of 16S rRNA gene primers for 454 pyrosequencing of the human foregut microbiome. *World J Gastroenterol*, **16**, 4135-4144.
33. Abed RM, A.-T.A., de Beer D. (2006) Bacterial diversity of a cyanobacterial mat degrading petroleum compounds at elevated salinities and temperatures. *Environ Microbiol*, **57**, 290-301.
34. Liebner S, H.J., Wagner D. (2008) Bacterial diversity and community structure in polygonal tundra soils from Samoylov Island, Lena Delta, Siberia. *Int Microbiol*, **11**, 195-202.
35. Schauer R, B.C., Ramette A, Harder J. (2010) Bacterial diversity and biogeography in deep-sea surface sediments of the South Atlantic Ocean. *ISME J*, **4**, 159-170.
36. Pruesse E, Q.C., Knittel K, Fuchs BM, Ludwig W, Peplies J, Glöckner FO. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acids Res*, **35**, 7188-7196.
37. DeSantis TZ, H.P., Larsen N, Rojas M, Brodie EL, Keller K, Huber T, Dalevi D, Hu P, Andersen GL. (2006) Greengenes, a chimera-checked 16S rRNA gene database and workbench compatible with ARB. *Appl Environ Microbiol*, **72**, 5069-5072.
38. Cole JR, C.B., Farris RJ, Wang Q, Kulam-Syed-Mohideen AS, McGarrell DM, Bandela AM, Cardenas E, Garrity GM, Tiedje JM. (2007) The ribosomal database

project (RDP-II): introducing myRDP space and quality controlled public data. *Nucleic Acids Res*, **35**, D169-D172.
39. Sanger F, N.S., Coulson AR,. (1977) DNA sequencing with chain-terminating inhibitors. . *Proc Natl Acad Sci U S A*, **74**, 5463-5467.
40. Harismendy O, N.P., Strausberg RL, Wang X, Stockwell TB, Beeson KY, Schork NJ, Murray SS, Topol EJ, Levy S, Frazer KA. (2009) Evaluation of next generation sequencing platforms for population targeted sequencing studies. . *Genome Biol*, **10**, R32.
41. Kircher M, K.J. (2010) High-throughput DNA sequencing – concepts and limitations. *BioEssays*, **32**, 524-536.
42. Tringe SG, H.P. (2008) A renaissance for the pioneering 16S rRNA gene. *Curr Opin Microbiol*, **11**, 442-446.
43. Margulies M, E.M., Altman WE, Attiya S, Bader JS, Bemben LA, Berka J, Braverman MS, Chen YJ, Chen Z, Dewell SB, Du L, Fierro JM, Gomes XV, Godwin BC, He W, Helgesen S, Ho CH, Irzyk GP, Jando SC, Alenquer ML, Jarvie TP, Jirage KB, Kim JB, Knight JR, Lanza JR, Leamon JH, Lefkowitz SM, Lei M, Li J, Lohman KL, Lu H, Makhijani VB, McDade KE, McKenna MP, Myers EW, Nickerson E, Nobile JR, Plant R, Puc BP, Ronan MT, Roth GT, Sarkis GJ, Simons JF, Simpson JW, Srinivasan M, Tartaro KR, Tomasz A, Vogt KA, Volkmer GA, Wang SH, Wang Y, Weiner MP, Yu P, Begley RF, Rothberg JM. (2005) Genome sequencing in microfabricated high-density picolitre reactors. *Nature*, **437**, 376-380.
44. Bennett S. (2004) Solexa Ltd. *Pharmacogenomics*, **5**, 433-438.
45. Liu L, L.Y., Li S, Hu N, He Y, Pong R, Lin D, Lu L, Law M. (2012) Comparison of Next-Generation Sequencing Systems. . *J Biomed Biotechnol*, **2012**, ID 251364.
46. Mardis ER. (2008) The impact of next-generation sequencing technology on genetics. *Trends Genet*, **24**, 133-141.
47. Valouev A, I.J., Tonthat T, Stuart J, Ranade S, Peckham H, Zeng K, Malek JA, Costa G, McKernan K, Sidow A, Fire A, Johnson SM. (2008) A high-resolution, nucleosome position map of C. elegans reveals a lack of universal sequence-dictated positioning. *Genome Res*, **18**, 1051-1063.
48. Rothberg JM, H.W., Rearick TM, Schultz J, Mileski W, Davey M, Leamon JH, Johnson K, Milgrew MJ, Edwards M, Hoon J, Simons JF, Marran D, Myers JW, Davidson JF, Branting A, Nobile JR, Puc BP, Light D, Clark TA, Huber M, Branciforte JT, Stoner IB, Cawley SE, Lyons M, Fu Y, Homer N, Sedova M, Miao X, Reed B, Sabina J, Feierstein E, Schorn M, Alanjary M, Dimalanta E, Dressman D, Kasinskas R, Sokolsky T, Fidanza JA, Namsaraev E, McKernan KJ, Williams A, Roth GT, Bustillo J. (2011) An integrated semiconductor device enabling non-optical genome sequencing. *Nature*, **475**, 348-352.
49. Liu Z, L.C., Hamady M, Bushman FD, Knight R,. (2007) Short pyrosequencing reads suffice for accurate microbial community analysis. *Nucleic Acids Res*, **35**, e120.
50. Gilbert JA, S.J., Caporaso JG, Steinbrück L, Reeder J, Temperton B, Huse S, McHardy AC, Knight R, Joint I, Somerfield P, Fuhrman JA, Field D. (2012) Defining seasonal marine microbial community dynamics. . *ISME J*, **6**, 289-308.
51. Caporaso JG, L.C., Walters WA, Berg-Lyons D, Huntley J, Fierer N, Owens SM, Betley J, Fraser L, Bauer M, Gormley N, Gilbert JA, Smith G, Knight R. (2012) Ultra-high-throughput microbial community analysis on the Illumina HiSeq and MiSeq platforms. *ISME J*, **6**, 1621-1624.
52. Caporaso JG, P.K., Field D, Knight R, Gilbert JA. (2012) The Western English Channel contains a persistent microbial seed bank. . *ISME J*, **6**, 1089-1093.

53. Bertrand H, P.F., Van VT, Lombard N, Nalin R, Vogel TM, Simonet P,. (2005) High molecular weight DNA recovery from soils prerequisite for biotechnological metagenomic library construction. *J Microbiol Methods*, **62**, 1-11.
54. Zhou HW, L.D., Tam NF, Jiang XT, Zhang H, Sheng HF, Qin J, Liu X, Zou F. (2011) BIPES, a cost-effective high-throughput method for assessing microbial diversity. *ISME J*, **5**, 741-749.
55. Lazarevic V, W.K., Huse S, Hernandez D, Farinelli L, Osterås M, Schrenzel J, François P. (2009) Metagenomic study of the oral microbiota by Illumina high-throughput sequencing. *J Microbiol Methods*, **79**, 266-271.
56. Claesson MJ, W.Q., O'Sullivan O, Greene-Diniz R, Cole JR, Ross RP, O'Toole PW. (2010) Comparison of two next-generation sequencing technologies for resolving highly complex microbiota composition using tandem variable 16S rRNA gene regions. *Nucleic Acids Res*, **38**, e200.
57. Gloor GB, H.R., Macklaim JM, Dickson RJ, Fernandes AD, MacPhee R, Reid G. (2010) Microbiome profiling by illumina sequencing of combinatorial sequence-tagged PCR products. *PLoS One*, **5**, e15406.
58. Whiteley AS, J.S., Waite I, Kresoje N, Payne H, Mullan B, Allcock R, O'Donnell A. (2012) Microbial 16S rRNA Ion Tag and community metagenome sequencing using the Ion Torrent (PGM) Platform. . *J Microbiol Methods*, **[Epub ahead of print]**.
59. Jünemann S, P.K., Szczepanowski R, Harks I, Ehmke B, Giesmann A, Stoye J, Harmsen D. (2012) Bacterial Community Shift in Treated Periodontitis Patients Revealed by Ion Torrent 16S rRNA Gene Amplicon Sequencing. . *PLoS One*, **7**, e41606.
60. Wommack KE, B.J., Ravel J. (2008) Metagenomics: read length matters. *Appl Environ Microbiol*, **74**, 1453-1463.
61. Chakravorty S, H.D., Burday M, Connell N, Alland D. (2007) A detailed analysis of 16S ribosomal RNA gene segments for the diagnosis of pathogenic bacteria. *J Microbiol Methods*, **69**, 330-339.
62. Eid J, F.A., Gray J, Luong K, Lyle J, Otto G, Peluso P, Rank D, Baybayan P, Bettman B, Bibillo A, Bjornson K, Chaudhuri B, Christians F, Cicero R, Clark S, Dalal R, Dewinter A, Dixon J, Foquet M, Gaertner A, Hardenbol P, Heiner C, Hester K, Holden D, Kearns G, Kong X, Kuse R, Lacroix Y, Lin S, Lundquist P, Ma C, Marks P, Maxham M, Murphy D, Park I, Pham T, Phillips M, Roy J, Sebra R, Shen G, Sorenson J, Tomaney A, Travers K, Trulson M, Vieceli J, Wegener J, Wu D, Yang A, Zaccarin D, Zhao P, Zhong F, Korlach J, Turner S. (2009) Real-time DNA sequencing from single polymerase molecules. *Science*, **323**, 133-138.
63. Armougom F, R.D. (2009) Exploring microbial diversity using 16S rRNA high-throughput methods. *J Comput Sci Syst Biol*, **2**, 74-92.
64. Rainey FA, W.-R.N., Janssen PH, Hippe H, Stackebrandt E. (1996) Clostridium paradoxum DSM 7308T contains multiple 16S rRNA genes with heterogeneous intervening sequences. *Microbiology*, **142**, 2087-2095.
65. Acinas SG, M.L., Klepac-Ceraj V, Pols MF. (2004) Divergence and redundancy of 16S rRNA sequences in genomes with multiple rrn operons. *J Bacteriol*, **186**, 2629-2635.
66. Stewart GC, W.F., Bott KF. (1982) Detailed physical mapping of the ribosomal RNA genes of Bacillus subtilis. *Gene*, **19**, 153-162.
67. Janda JM, A.S. (2007) 16S rRNA gene sequencing for bacterial identification in the diagnostic laboratory: pluses, perils, and pitfalls. *J Clin Microbiol*, **45**, 2761-2764.
68. Adékambi T, C.P., Drancourt M. (2003) rpoB-based identification of nonpigmented and late-pigmenting rapidly growing mycobacteria. *J Clin Microbiol*, **41**, 5699-5708.

69. Baker GC, S.J., Cowan DA. (2003) Review and re-analysis of domain-specific 16S primers. *J Microbiol Methods*, **55**, 541-555.
70. Wang Y, Q.P. (2009) Conservative Fragments in Bacterial 16S rRNA Genes and Primer Design for 16S Ribosomal DNA Amplicons in Metagenomic Studies *PLoS One*, **4**, e7401.
71. Andersson AF, L.M., Jakobsson H, Bäckhed F, Nyrén P, Engstrand L. (2008) Comparative analysis of human gut microbiota by barcoded pyrosequencing. *PLoS One*, **3**, e2836.
72. Medini D, S.D., Parkhill J, Relman DA, Donati C, Moxon R, Falkow S, Rappuoli R. (2008) Microbiology in the post-genomic era. *Nat Rev Microbiol*, **6**, 419-430.
73. Leininger S, U.T., Schloter M, Schwark L, Qi J, Nicol GW, Prosser JI, Schuster SC, Schleper C. (2006) Archaea predominate among ammonia-oxidizing prokaryotes in soils. *Nature*, **442**, 806-809.
74. Urich T, L.A., Qi J, Huson DH, Schleper C, Schuster SC. (2008) Simultaneous assessment of soil microbial community structure and function through analysis of the meta-transcriptome. *PLoS One*, **3**, e2527.
75. Frias-Lopez J, S.Y., Tyson GW, Coleman ML, Schuster SC, Chisholm SW, Delong EF. (2008) Microbial community gene expression in ocean surface waters. *Proc Natl Acad Sci U S A*, **105**, 3805-3810.
76. Gilbert JA, F.D., Huang Y, Edwards R, Li W, Gilna P, Joint I. (2008) Detection of Large Numbers of Novel Sequences in the Metatranscriptomes of Complex Marine Microbial Communities. *PLoS One*, **3**, e3042.
77. Poretsky RS, H.I., Sun S, Allen AE, Zehr JP, Moran MA. (2009) Comparative day/night metatranscriptomic analysis of microbial communities in the North Pacific subtropical gyre. *Environ Microbiol*, **11**, 1358-1375.
78. Helbling DE, A.M., Fenner K, Kohler HP, Johnson DR. (2012) The activity level of a microbial community function can be predicted from its metatranscriptome. *ISME J*, **6**, 902-904.
79. Bomar L, M.M., Colston S, Graf J. (2011) Directed culturing of microorganisms using metatranscriptomics. *Mbio*, **5**, e00012-00011.
80. Liu Z, K.C., Wood JM, Rusch DB, Ludwig M, Wittekindt N, Tomsho LP, Schuster SC, Ward DM, Bryant DA. (2011) Metatranscriptomic analyses of chlorophototrophs of a hot-spring microbial mat. . *ISME J*, **5**, 1279-1290.
81. Vila-Costa M, R.-K.J., Sun S, Sharma S, Poretsky R, Moran MA. (2010) Transcriptomic analysis of a marine bacterial community enriched with dimethylsulfoniopropionate. *ISME J*, **4**, 1410-1420.
82. Wu J, G.W., Zhang W, Meldrum DR. (2011) Optimization of whole-transcriptome amplification from low cell density deep-sea microbial samples for metatranscriptomic analysis. *J Microbiol Methods*, **84**, 88-93.
83. Stewart FJ, U.O., DeLong EF. (2012) Microbial metatranscriptomics in a permanent marine oxygen minimum zone. *Environ Microbiol*, **14**, 23-40.
84. Rinta-Kanto JM, S.S., Sharma S, Kiene RP, Moran MA. (2012) Bacterial community transcription patterns during a marine phytoplankton bloom. *14*, **1**.
85. Gosalbes MJ, D.A., Pignatelli M, Abellan JJ, Jiménez-Hernández N, Pérez-Cobas AE, Latorre A, Moya A. (2011) Metatranscriptomic approach to analyze the functional human gut microbiota. *PLoS One*, **6**, e17447.
86. de Menezes A, C.N., Doyle E. (2012) Comparative metatranscriptomics reveals widespread community responses during phenanthrene degradation in soil. . *Environ Microbiol*, [Epub ahead of print].
87. Croucher NJ, T.N. (2010) Studying bacterial transcriptomes using RNA-seq. *Curr Opin Microbiol*, **13**, 619-624.

88. Filiatrault MJ. (2011) Progress in prokaryotic transcriptomics. *Curr Opin Microbiol*, **14**, 579-586.
89. Shi Y, T.G., DeLong EF. (2009) Metatranscriptomics reveals unique microbial small RNAs in the ocean's water column. *Nature*, **459**, 266-269.
90. Sittka A, L.S., Papenfort K, Sharma CM, Rolle K, Binnewies TT, Hinton JC, Vogel J,. (2008) Deep sequencing analysis of small noncoding RNARNA and mRNARNA targets of the global post-transcriptional regulator, Hfq. . *PLoS Genet*, **4**, e1000163.
91. Warnecke F, H.M. (2009) A perspective: metatranscriptomics as a tool for the discovery of novel biocatalysts. *J Biotechnol*, **142**, 91-95.
92. Brenner S, J.F., Meselson M. (1961) An unstable intermediate carrying information from genes to ribosomes for protein synthesis. *Nature*, **190**, 576-581.
93. Gros F, H.H., Gilbert W, Kurland CG, Risebrough RW, Watson JD. (1961) Unstable ribonucleic acid revealed by pulse labelling of Escherichia coli+. *Nature*, **190**, 581-585.
94. Steege DA. (2000) Emerging features of mRNA decay in bacteria. *RNA*, **6**, 1079-1090.
95. Redon E, L.P., Cocaign-Bousquet M. (2005) Role of mRNA stability during genome-wide adaptation of Lactococcus lactis to carbon starvation. . *J Biol Chem*, **280**, 36380-36385.
96. Sharova LV, S.A., Nedorezov T, Piao Y, Shaik N, Ko MSH, . (2009) Database for mRNA Half-Life of 19 977 Genes Obtained by DNA Microarray Analysis of Pluripotent and Differentiating Mouse Embryonic Stem Cells. *DNA Res*, **16**, 45-58.
97. Selinger DW, S.R., Cheung KJ, Church GM, Rosenow C. (2003) Global RNA Half-Life Analysis in Escherichia coli Reveals Positional Patterns of Transcript Degradation. *Genome Res*, **13**, 216-223.
98. Kaberdin VR, B.U. (2006) Translation initiation and the fate of bacterial mRNAs. *FEMS Microbiol Rev*, **30**, 967-979.
99. Sorek R, C.P. (2010) Prokaryotic transcriptomics: a new view on regulation, physiology and pathogenicity. *Nat Rev Genet*, **11**, 9-16.
100. Rio DC, A.M.J., Hannon GJ, Nilsen TW. (2012) Enrichment of Poly(A)+ mRNA Using Immobilized Oligo(dT). . *CSHLP*, **2010**.
101. Stewart FJ, O.E., DeLong EF. (2010) Development and quantitative analyses of a universal rRNA-subtraction protocol for microbial metatranscriptomics. . *ISME J*, **4**, 896-907.
102. Pang X, Z.D., Song Y, Pei D, Wang J, Guo Z, Yang R. (2004) Bacterial mRNA purification by magnetic capture-hybridization method. *Microbiol Immunol*, **48**, 91-96.
103. Dunman PM, M.E., Haney S, Palacios D, Tucker-Kellogg G, Wu S, Brown EL, Zagursky RJ, Shlaes D, Projan SJ. (2001) Transcription profiling-based identification of Staphylococcus aureus genes regulated by the agr and/or sarA loci. *J Bacteriol*, **183**, 7341-7353.
104. McGrath KC, T.-H.S., Cheng CT, Leo L, Alexa A, Schmidt S, Schenk PM. (2008) Isolation and analysis of mRNA from environmental microbial communities. *J Microbiol Methods*, **75**, 172-176.
105. Mettel C, K.Y., Shrestha PM, Liesack W. (2010) Extraction of mRNA from Soil. . *Appl Environ Microbiol*, **76**, 5995-6000.
106. Croucher NJ, F.M., Perkins TT, Turner DJ, Marguerat SB, Keane T, Quail MA, He M, Assefa S, Bähler J, Kingsley RA, Parkhill J, Bentley SD, Dougan G, Thomson NR. (2009) A simple method for directional transcriptome sequencing using Illumina technology. *Nucleic Acids Res*, **37**, e418.

107. Hansen KD, B.S., Dudoit S. (2010) Biases in Illumina transcriptome sequencing caused by random hexamer priming. *Nucleic Acids Res*, **38**, e131.
108. Weber AP, W.K., Carr K,Wilkerson C, Ohlrogge JB. (2007) Sampling the Arabidopsis transcriptome with massively parallel pyrosequencing. *Plant Physiol*, **144**, 32-42.
109. Cheung F, H.B., Goldberg SM, May GD, Xiao Y, Town CD. . (2006) Sequencing Medicago truncatula expressed sequenced tags using 454 Life Sciences technology. . *BMC Genomics*, **7**, 272.
110. Holmes C, M.F., Jones M, Ozdemir V, Graham JE. (2010) Standardization and Omics Science: Technical and Social Dimensions Are Inseparable and Demand Symmetrical Study. . *OMICS*, **14**, 327-332.
111. van Vliet AH. (2010) Next generation sequencing of microbial transcriptomes: challenges and opportunities. *FEMS Microbiol Lett*, **302**, 1-7.
112. Brazma A, H.P., Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, Gaasterland T, Glenisson P, Holstege FC, Kim IF, Markowitz V, Matese JC, Parkinson H, Robinson A, Sarkans U, Schulze-Kremer S, Stewart J, Taylor R, Vilo J, Vingron M. (2001) Minimum information about a microarray experiment (MIAME)-toward standards for microarray data. *Nat Genet*, **29**, 365-371.
113. Field D, G.G., Gray T, Morrison N, Selengut J, Sterk P, Tatusova T, Thomson N, Allen MJ, Angiuoli SV, Ashburner M, Axelrod N, Baldauf S, Ballard S, Boore J, Cochrane G, Cole J, Dawyndt P, De Vos P, DePamphilis C, Edwards R, Faruque N, Feldman R, Gilbert J, Gilna P, Glöckner FO, Goldstein P, Guralnick R, Haft D, Hancock D, Hermjakob H, Hertz-Fowler C, Hugenholtz P, Joint I, Kagan L, Kane M, Kennedy J, Kowalchuk G, Kottmann R, Kolker E, Kravitz S, Kyrpides N, Leebens-Mack J, Lewis SE, Li K, Lister AL, Lord P, Maltsev N, Markowitz V, Martiny J, Methe B, Mizrachi I, Moxon R, Nelson K, Parkhill J, Proctor L, White O, Sansone SA, Spiers A, Stevens R, Swift P, Taylor C, Tateno Y, Tett A, Turner S, Ussery D, Vaughan B, Ward N, Whetzel T, San Gil I, Wilson G, Wipat A. (2008) The minimum information about a genome sequence (MIGS) specification. . *Nat Biotechnol*, **26**, 541-547.
114. Yilmaz P, K.R., Field D, Knight R, Cole JR, Amaral-Zettler L, Gilbert JA, Karsch-Mizrachi I, Johnston A, Cochrane G, Vaughan R, Hunter C, Park J, Morrison N, Rocca-Serra P, Sterk P, Arumugam M, Bailey M, Baumgartner L, Birren BW, Blaser MJ, Bonazzi V, Booth T, Bork P, Bushman FD, Buttigieg PL, Chain PS, Charlson E, Costello EK, Huot-Creasy H, Dawyndt P, DeSantis T, Fierer N, Fuhrman JA, Gallery RE, Gevers D, Gibbs RA, San Gil I, Gonzalez A, Gordon JI, Guralnick R, Hankeln W, Highlander S, Hugenholtz P, Jansson J, Kau AL, Kelley ST, Kennedy J, Knights D, Koren O, Kuczynski J, Kyrpides N, Larsen R, Lauber CL, Legg T, Ley RE, Lozupone CA, Ludwig W, Lyons D, Maguire E, Methé BA, Meyer F, Muegge B, Nakielny S, Nelson KE, Nemergut D, Neufeld JD, Newbold LK, Oliver AE, Pace NR, Palanisamy G, Peplies J, Petrosino J, Proctor L, Pruesse E, Quast C, Raes J, Ratnasingham S, Ravel J, Relman DA, Assunta-Sansone S, Schloss PD, Schriml L, Sinha R, Smith MI, Sodergren E, Spo A, Stombaugh J, Tiedje JM, Ward DV, Weinstock GM, Wendel D, White O, Whiteley A, Wilke A, Wortman JR, Yatsunenko T, Glöckner FO. (2011) Minimum information about a marker gene sequence (MIMARKS) and minimum information about any (x) sequence (MIxS) specifications. *Nat Biotechnol*, **29**, 415-420.
115. Nie L, W.G., Culley DE, Scholten JC, Zhang W. (2007) Integrative Analysis of Transcriptomic and Proteomic Data: Challenges, Solutions and Applications. . *Crit Rev Biotechnol*, **27**, 63-75.

116. Zhang W, L.F., Nie L. (2010) Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, **156**, 287-301.
117. Siggins A, G.E., Abram F. (2012) Exploring mixed microbial community functioning: recent advances in metaproteomics. *Environ Microbiol*, **80**, 265-280.
118. Venter JC, A.M., Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA, Gocayne JD, Amanatides P, Ballew RM, Huson DH, Wortman JR, Zhang Q, Kodira CD, Zheng XH, Chen L, Skupski M, Subramanian G, Thomas PD, Zhang J, Gabor Miklos GL, Nelson C, Broder S, Clark AG, Nadeau J, McKusick VA, Zinder N, Levine AJ, Roberts RJ, Simon M, Slayman C, Hunkapiller M, Bolanos R, Delcher A, Dew I, Fasulo D, Flanigan M, Florea L, Halpern A, Hannenhalli S, Kravitz S, Levy S, Mobarry C, Reinert K, Remington K, Abu-Threideh J, Beasley E, Biddick K, Bonazzi V, Brandon R, Cargill M, Chandramouliswaran I, Charlab R, Chaturvedi K, Deng Z, Di Francesco V, Dunn P, Eilbeck K, Evangelista C, Gabrielian AE, Gan W, Ge W, Gong F, Gu Z, Guan P, Heiman TJ, Higgins ME, Ji RR, Ke Z, Ketchum KA, Lai Z, Lei Y, Li Z, Li J, Liang Y, Lin X, Lu F, Merkulov GV, Milshina N, Moore HM, Naik AK, Narayan VA, Neelam B, Nusskern D, Rusch DB, Salzberg S, Shao W, Shue B, Sun J, Wang Z, Wang A, Wang X, Wang J, Wei M, Wides R, Xiao C, Yan C, Yao A, Ye J, Zhan M, Zhang W, Zhang H, Zhao Q, Zheng L, Zhong F, Zhong W, Zhu S, Zhao S, Gilbert D, Baumhueter S, Spier G, Carter C, Cravchik A, Woodage T, Ali F, An H, Awe A, Baldwin D, Baden H, Barnstead M, Barrow I, Beeson K, Busam D, Carver A, Center A, Cheng ML, Curry L, Danaher S, Davenport L, Desilets R, Dietz S, Dodson K, Doup L, Ferriera S, Garg N, Gluecksmann A, Hart B, Haynes J, Haynes C, Heiner C, Hladun S, Hostin D, Houck J, Howland T, Ibegwam C, Johnson J, Kalush F, Kline L, Koduru S, Love A, Mann F, May D, McCawley S, McIntosh T, McMullen I, Moy M, Moy L, Murphy B, Nelson K, Pfannkoch C, Pratts E, Puri V, Qureshi H, Reardon M, Rodriguez R, Rogers YH, Romblad D, Ruhfel B, Scott R, Sitter C, Smallwood M, Stewart E, Strong R, Suh E, Thomas R, Tint NN, Tse S, Vech C, Wang G, Wetter J, Williams S, Williams M, Windsor S, Winn-Deen E, Wolfe K, Zaveri J, Zaveri K, Abril JF, Guigó R, Campbell MJ, Sjolander KV, Karlak B, Kejariwal A, Mi H, Lazareva B, Hatton T, Narechania A, Diemer K, Muruganujan A, Guo N, Sato S, Bafna V, Istrail S, Lippert R, Schwartz R, Walenz B, Yooseph S, Allen D, Basu A, Baxendale J, Blick L, Caminha M, Carnes-Stine J, Caulk P, Chiang YH, Coyne M, Dahlke C, Mays A, Dombroski M, Donnelly M, Ely D, Esparham S, Fosler C, Gire H, Glanowski S, Glasser K, Glodek A, Gorokhov M, Graham K, Gropman B, Harris M, Heil J, Henderson S, Hoover J, Jennings D, Jordan C, Jordan J, Kasha J, Kagan L, Kraft C, Levitsky A, Lewis M, Liu X, Lopez J, Ma D, Majoros W, McDaniel J, Murphy S, Newman M, Nguyen T, Nguyen N, Nodell M, Pan S, Peck J, Peterson M, Rowe W, Sanders R, Scott J, Simpson M, Smith T, Sprague A, Stockwell T, Turner R, Venter E, Wang M, Wen M, Wu D, Wu M, Xia A, Zandieh A, Zhu X. (2001) The Sequence of the Human Genome. . *Science*, **291**, 1304-1351.
119. Venter JC, R.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 YH, Smith HO. (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 554-557.
120. Metzker M. (2010) Sequencing technologies — the next generation. *Nat Rev Genet*, **11**, 31-45.
121. Metzker ML. (2005) Emerging technologies in DNA sequencing. . *Genome Res*, **15**, 1767-1776.

122. Pozhitkov AE, B.T., Flemmig T, Noble PA. (2011) High-throughput methods for analysis of the human oral microbiome. *Periodontology 2000*, **55**, 70-86.
123. Oyola SO, O.T., Gu Y, Maslen G, Manske M, Campino S, Turner DJ, Macinnis B, Kwiatkowski DP, Swerdlow HP, Quail MA. (2012) Optimizing illumina next-generation sequencing library preparation for extremely at-biased genomes. . *BMC Genomics*, **13**.
124. Loman NJ, M.R., Dallman TJ, Constantinidou C, Gharbia SE, Wain J, Pallen MJ. (2012) Performance comparison of benchtop high-throughput sequencing platforms. *Nat Biotechnol*, **30**, 434-439.
125. Koren S, S.M., Walenz BP, Martin J, Howard JT, Ganapathy G, Wang Z, Rasko DA, McCombie WR, Jarvis ED, Phillippy AM. (2012) Hybrid error correction and de novo assembly of single-molecule sequencing reads. . *Nat Biotechnol*, **doi: 10.1038/nbt.2280**, [Epub ahead of print].
126. Wecker P, K.C., Ellrott A, Quast C, Langhammer P, Harder J and Glöckner FO. (2009) Transcriptional response of the model planctomycete Rhodopirellula baltica SH1T to changing environmental conditions. *BMC Genomics*, **10**.
127. Franke HD, B.F., Wiltshire KH. (2004) Ecological long-term research at Helgoland (German Bight, North Sea): retrospect and prospect—an introduction. *Helgol Mar Res*, **58**, 223-229.
128. Gerdts G, W.A., Döpke H, Klings KW, Gunkel W, Schütt C. (2004) 40-year long-term study of microbial parameters near Helgoland (German Bight, North Sea): historical view and future perspectives. *Helgol Mar Res*, **58**, 230-242.
129. Schlesner H, R.C., Tindall BJ, Gade D, Rabus R, Pfeiffer S, Hirsch P. (2004) Taxonomic heterogeneity within the Planctomycetales as derived by DNA–DNA hybridization, description of Rhodopirellula baltica gen. nov., sp. nov., transfer of Pirellula marina to the genus Blastopirellula gen. nov. as Blastopirellula marina comb. nov. and emended description of the genus Pirellula. *Int J Syst Evol Microbiol*, **54**, 1567-1580.
130. Schlesner H. (1994) The development of media suitable for the microorganisms morphologically resembling Planctomyces spp., Pirellula spp., and other Planctomycetales from various aquetic habitats using dilute media. *System Appl Microbiol*, **17**, 135-145.
131. DeLong EF, F.D., Alldredge AI. (1993) Phylogenetic diversity of aggregate-attached vs. free-living marine bacterial assemblages. *Limnol Oceanogr*, **38**, 924-934.
132. Gade D, T.D., Lange D, Mirgorodskaya E, Lombardot T, Glöckner FO, Kube M, Reinhardt R, Amann R, Lehrach H, Rabus R, Gobom J. (2005) Towards the proteome of the marine bacterium Rhodopirellula baltica: mapping the soluble proteins. *Proteomics*, **5**, 3654-3671.
133. König H, S.H., Hirsch P. (1984) Cell wall studies on budding bacteria of the Planctomyces/Pasteuria group and on a Prosthecomicrobium sp. *Arch Microbiol*, **138**, 200–205.
134. Liesack W, K.H., Schlesner H, Hirsch P. (1986) Chemical composition of the peptidoglycan-free cell envelopes of budding bacteria of the Pirella/Planctomyces group. *Arch Microbiol*, **145**, 361–366.
135. Lindsay M, W.R., Strous M, Jetten MSM, Butler MK, Forde RJ, Fuerst JA. (2001) Cell compartmentalisation in planctomycetes: novel types of structural organisation for the bacterial cell. *Arch Microbiol*, **175**, 413-429.
136. Gade D, S.T., Reinhardt R, Rabus R. (2005) Growth phase dependent regulation of protein composition in Rhodopirellula baltica. *Environ Microbiol*, **7**, 1074-1084.
137. Schloss PD, G.D., Westcott SL. (2011) Reducing the Effects of PCR Amplification and Sequencing Artifacts on 16S rRNA-Based Studies. *PLoS One*, **6**, e27310.

138. Liu Z, D.T., Andersen GL, Knight R. (2008) Accurate taxonomy assignments from 16S rRNA sequences produced by highly parallel pyrosequencers. *Nucleic Acids Res*, **36**, e120.
139. Alm EW, O.D., Larsen N, Stahl DA, Raskin L. (1996) The oligonucleotide probe database. *Appl Environ Microbiol*, **62**, 3557-3559.
140. Teeling H, F.B., Becher D, Klockow C, Gardebrecht A, Bennke CM, Kassabgy M, Huang S, Mann AJ, Waldmann J, Weber M, Klindworth A, Otto A, Lange J, Bernhardt J, Reinsch C, Hecker M, Peplies J, Bockelmann FD, Callies U, Gerdts G, Wichels A, Wiltshire KH, Glöckner FO, Schweder T, Amann R. (2012) Substrate-controlled succession of marine bacterioplankton populations induced by a phytoplankton bloom. *Science*, **336**, 608-611.
141. Brosius J, P.M., Kennedy PJ, Noller HF. (1978) Complete nucleotide sequence of a 16S ribosomal RNA gene from Escherichia coli. *Proc Natl Acad Sci U S A*, **75**, 4801-4805.
142. Kibbe WA. (2007) OligoCalc: an online oligonucleotide properties calculator. *Nucleic Acids Res*, **35**, W43-46.
143. Pruesse E, P.J., Glöckner FO. (2012) SINA: accurate high throughput multiple sequence alignment of ribosomal RNA genes. *Bioinformatics*, **28**, 1823-1829.
144. Edgar RC. (2010) Search and clustering orders of magnitude faster than BLAST. *Bioinformatics*, **26**, 2460-2461.
145. Rusch DB, H.A., Sutton G, Heidelberg KB, Williamson S, Yooseph S, Wu D, Eisen JA, Hoffman JM, Remington K, Beeson K, Tran B, Smith H, Baden-Tillson H, Stewart C, Thorpe J, Freeman J, Andrews-Pfannkoch C, Venter JE, Li K, Kravitz S, Heidelberg JF, Utterback T, Rogers YH, Falcón LI, Souza V, Bonilla-Rosso G, Eguiarte LE, Karl DM, Sathyendranath S, Platt T, Bermingham E, Gallardo V, Tamayo-Castillo G, Ferrari MR, Strausberg RL, Nealson K, Friedman R, Frazier M, Venter JC. (2007) The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS One*, **5**, e77.
146. Yooseph S, S.G., Rusch DB, Halpern AL, Williamson SJ, Remington K, Eisen JA, Heidelberg KB, Manning G, Li W, Jaroszewski L, Cieplak P, Miller CS, Li H, Mashiyama ST, Joachimiak MP, van Belle C, Chandonia JM, Soergel DA, Zhai Y, Natarajan K, Lee S, Raphael BJ, Bafna V, Friedman R, Brenner SE, Godzik A, Eisenberg D, Dixon JE, Taylor SS, Strausberg RL, Frazier M, Venter JC. (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol*, **5**, e16.
147. Ludwig W, S.O., Westram R, Richter L, Meier H, Yadhukumar, Buchner A, Lai T, Steppi S, Jobb G, Förster W, Brettske I, Gerber S, Ginhart AW, Gross O, Grumann S, Hermann S, Jost R, König A, Liss T, Lüssmann R, May M, Nonhoff B, Reichel B, Strehlow R, Stamatakis A, Stuckmann N, Vilbig A, Lenke M, Ludwig T, Bode A, Schleifer KH. (2004) ARB: a software environment for sequence data. . *Nucleic Acids Res*, **32**, 1363-1371.
148. Zhou J, B.M., Tiedje JM. (1996) DNA recovery from soils of diverse composition. *Appl Environ Microbiol*, **62**, 316-322.
149. Herlemann DPR, L.M., Juergens K, Bertilsson S, Waniek JJ, Anderrson AF. (2011) Transition in bacterial communities along the 2000 km salinity gradient of the Baltic Sea. *ISME J*, **5**, 1571-1579.
150. Muyzer G, d.W.E., Uitterlinden AG. (1993) Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16S rRNA. *Appl Environ Microbiol*, **59**, 695–700.
151. Muyzer G, B.T., Nuebel U, Santegoeds C, Schaefer H, Waver C. (1998) Denaturing gradient gel electrophoresis (DGGE) in microbial ecology. *In: Akkermans ADL, van*

Elsas JD, de Bruijn FJ (eds). Molecular Microbial Ecology Manual. Kluwer Academic Publishers: Dordrecht, The Netherlands., 1-27.
152. Gilbert JA, F.D., Swift P, Thomas S, Cummings D, Temperton B, Weynberg K, Huse S, Hughes M, Joint I, Somerfield PJ, Mühling M. (2010) The taxonomic and functional diversity of microbes at a temperate coastal site: a 'multi-omic' study of seasonal and diel temporal variation. *PLoS One*, **5**, e15545.
153. Walters WA, C.J., Lauber CL,Berg-Lyons D, Fierer N, Knight R. (2011) PrimerProspector: de novo design and taxonomic analysis of barcoded PCR primers. *Bioinformatics*, **27**, 1159-1161.
154. Barns SM, F.R., Jeffries MW, Pace NR. (1994) Remarkable archaeal diversity detected in a Yellowstone National Park hot spring environment. *Proc Natl Acad Sci U S A*, **91**, 1609-1613.
155. Reysenbach AL, P.N. (1995) In: Robb, F.T., Place, A.R. (Eds.), Archaea: A Laboratory Manual—Thermophiles. . *CSHLP*, 101-107.
156. Huws SA, E.J., Kim EJ, Scollan ND. (2007) Specificity and sensitivity of eubacterial primers utilized for molecular profiling of bacteria within complex microbial ecosystems. *J Microbiol Methods*, **70**, 565-569.
157. Droege M, H.B. (2008) The Genome Sequencer FLX System--longer reads, more applications, straight forward bioinformatics and more complete data sets. *J Biotechnol*, **136**, 3-10.
158. Sipos R, S.A., Palatinszky M, Révész S, Márialigeti K, Nikolausz M. (2007) Effect of primer mismatch, annealing temperature and PCR cycle number on 16S rRNA gene-targetting bacterial community analysis. *Environ Microbiol*, **60**, 341-350.
159. Baker GC, C.D. (2004) 16 S rDNA primers and the unbiased assessment of thermophile diversity. *Biochem Soc Trans*, **32**, 218-221.
160. Huber H, H.M., Rachel R, Fuchs T, Wimmer VC, Stetter KO. (2002) A new phylum of Archaea represented by a nanosized hyperthermophilic symbiont. *Nature*, **417**, 63-67.
161. McCliment EA, N.C., Carlson CA, Alldredge AL, Witting J, Amaral-Zettler LA. (2012) An all-taxon microbial inventory of the Moorea coral reef ecosystem. *ISME J*, **6**, 309-319.
162. Yooseph S, N.K., Rusch DB, McCrow JP, Dupont CL, Kim M, Johnson J, Montgomery R, Ferriera S, Beeson K, Williamson SJ, Tovchigrechko A, Allen AE, Zeigler LA, Sutton G, Eisenstadt E, Rogers YH, Friedman R, Frazier M, Venter JC. (2010) Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, **468**, 60-66.
163. Gilbert JA, M.F., Jansson J, Gordon J, Pace N, Tiedje J, Ley R, Fierer N, Field D, Kyrpides N, Glöckner FO, Klenk HP, Wommack KE, Glass E, Docherty K, Gallery R, Stevens R, Knight R. (2010) The Earth Microbiome Project: Meeting report of the "1st EMP meeting on sample selection and acquisition" at Argonne National Laboratory October 6th 2010. *Stand Genomic Sci*, **3**, 249-253.
164. Gattuso JP, F.M., Wollastr R. (1998) Carbon and carbonate metabolism in coastal aquatic ecosystems. *Annu Rev Ecol Syst*, **29**, 405-434.
165. Field CB, B.M., Randerson JT, Falkowski P. (1998) Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, **281**, 237-240.
166. Azam F. (1998) Microbial control of oceanic carbon flux: The plot thickens. . *Science*, **280**.
167. Azam F, F.T., Field JG, Gray JS, Meyer-Reil LA, Thingstad F. (1983) The ecological role of water-column microbes in the sea. . *Mar Ecol Prog Ser*, **10**, 257-263.

168. Pinhassi J, A.F., Hemphala J, Long RA, Martinez J, Zweifel UL, Hagström (1999) Coupling between bacterioplankton species composition, population dynamics, and organic matter degradation. . *Aquat Microbiol Ecol*, **17**, 13-26.
169. Riemann L, S.G., Azam F. (2000) Dynamics of bacterial community composition and activity during a mesocosm diatom bloom. . *Appl Environ Microbiol*, **66**, 578-587.
170. Pinhassi J, S.M., Havskum H, Peters F, Guadayol O, Malits A, Marrasé C. (2004) Changes in Bacterioplankton Composition under Different Phytoplankton Regimens. . *Appl Environ Microbiol*, **70**, 6753-6766.
171. Fandino LB, R.L., Steward GF, Long RA, Azam F. (2001) Variations in bacterial community structure during a dinoflagellate bloom analyzed by DGGE and 16S rDNA sequencing. . *Aquat Microbiol Ecol*, **23**, 119-130.
172. Lau WW, K.R., Armbrust EV. (2007) Succession and diel transcriptional response of the glycolate-utilizing component of the bacterial community during a spring phytoplankton bloom. . *Appl Environ Microbiol*, **73**, 2440-2450.
173. Tada Y, T.A., Nagao I, Miki T, Uematsu M, Tsuda A, Hamasaki K. (2011) Differing growth responses of major phylogenetic groups of marine bacteria to natural phytoplankton blooms in the western North Pacific Ocean. . *Appl Environ Microbiol*, **77**, 4055-4065.
174. Cantarel BL, C.P., Rancurel C, Bernard T, Lombard V, Henrissat B. (2009) The Carbohydrate-Active EnZymes database (CAZy): an expert resource for Glycogenomics. *Nucleic Acids Res*, **37**, D233-238.
175. Wustman BA, G.M., Hoagland KD. (1997) Extracellular Matrix Assembly in Diatoms (Bacillariophyceae) (I. A Model of Adhesives Based on Chemical Characterization and Localization of Polysaccharides from the Marine Diatom Achnanthes longipes and Other Diatoms). *Plant Physiol*, **113**, 1059-1069.
176. Khodse VB, B.N. (2010) Differences in carbohydrate profiles in batch culture grown planktonic and biofilm cells of Amphora rostrata Wm. Sm. *Biofouling*, **26**, 527-537.
177. Gómez-Pereira PR, F.B., Alonso C, Oliver MJ, van Beusekom JE, Amann R. (2010) Distinct flavobacterial communities in contrasting water masses of the north Atlantic Ocean. *ISME J*, **4**, 472-487.
178. Braun V, H.K. (2011) Recent insights into iron import by bacteria. . *Curr Opin Chem Biol*, **15**, 328-334.
179. Krewulak KD, V.H. (2011) TonB or not TonB: Is that the question? . *Biochem Cell Biol*, **89**, 87-97.
180. Rostovtseva TK, N.E., Bezrukov SM. (2002) Partitioning of differently sized poly(ethylene glycol)s into OmpF porin. . *Biophys J*, **82**, 160-169.
181. Thomas F, H.J., Rebuffet E, Czjzek M, Michel G. (2011) Environmental and gut bacteroidetes: The food connection. . *Front Microbiol.*, **2**, 93.
182. Hopkinson BM, B.K. (2012) Iron transporters in marine prokaryotic genomes and metagenomes. . *Environ Microbiol*, **14**, 114-128.
183. Bauer M, K.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. (2006) Whole genome analysis of the marine Bacteroidetes'Gramella forsetii' reveals adaptations to degradation of polymeric organic matter. *Environ Microbiol*, **8**, 2201-2213.
184. Schauer K, R.D., de Reuse H. (2008) New substrates for TonB-dependent transport: do we only see the 'tip of the iceberg'? *Trends Biochem Sci*, **33**, 330-338.
185. Morris RM, N.B., Frazar C, Goodlett DR, Ting YS, Rocap G. (2010) Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *ISME J*, **4**, 673-675.

186. Giovannoni SJ, T.H., Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ. (2005) Genome streamlining in a cosmopolitan oceanic bacterium. *Science*, **309**, 1242-1245.
187. Reisch CR, S.M., Varaljay VA, Amster IJ, Moran MA, Whitman WB. (2011) Novel pathway for assimilation of dimethylsulphoniopropionate widespread in marine bacteria. . *Nature*, **473**, 208-211.
188. Sun M, Y.Y., Yang P, Lei B, Du L, Kijlstra A. (2011) Regulatory effects of IFN-β on the development of experimental autoimmune uveoretinitis in B10RIII mice. . *PLoS One*, **6**, e19870.
189. Sowell SM, W.L., Norbeck AD, Lipton MS, Nicora CD, Barofsky DF, Carlson CA, Smith RD, Giovanonni SJ. (2009) Transport functions dominate the SAR11 metaproteome at low-nutrient extremes in the Sargasso Sea. . *ISME J*, **3**, 93-105.
190. Moran MA, B.R., Schell MA, González JM, Sun F, Sun S, Binder BJ, Edmonds J, Ye W, Orcutt B, Howard EC, Meile C, Palefsky W, Goesmann A, Ren Q, Paulsen I, Ulrich LE, Thompson LS, Saunders E, Buchan A. (2007) Ecological Genomics of Marine Roseobacters. . *Appl Environ Microbiol*, **73**, 4559-4569.
191. Brinkhoff T, G.H., Simon M. (2008) Diversity, ecology, and genomics of the Roseobacter clade: A short overview. . *Arch Microbiol*, **189**, 531-539.
192. Newton RJ, G.L., Bowles KM, Meile C, Gifford S, Givens CE, Howard EC, King E, Oakley CA, Reisch CR, Rinta-Kanto JM, Sharma S, Sun S, Varaljay V, Vila-Costa M, Westrich JR, Moran MA. (2010) Genome characteristics of a generalist marine bacterial lineage. . *ISME J*, **4**, 784-798.
193. Tittel J, B.O., Kamjunke N. (2011) Non-cooperative behaviour of bacteria prevents efficient phosphorus utilization of planktonic communities. . *J Plankton Res*, **34**, 102-112.
194. Woebken D, F.B., Kuypers MM, Amann R. (2007) Potential interactions of particle-associated anammox bacteria with bacterial and archaeal partners in the Namibian upwelling system. . *Appl Environ Microbiol*, **73**, 4648-4657.
195. Edwards JL, S.D., Connolly J, McDonald JE, Cox MJ, Joint I, Edwards C, McCarthy AJ,. (2010) Identification of carbohydrate metabolism genes in the metagenome of a marine biofilm community shown to be dominated by gammaproteobacteria and bacteroidetes. . *Genes*, **1**, 371-384.
196. Arnosti C. (2011) Microbial extracellular enzymes and the marine carbon cycle. . *Ann Rev Mar Sci*, **3**, 401-425.
197. Romanenko LA, S.P., Rohde M, Mikhailov VV, Stackebrandt E. (2004) "Reinekea marinisedimentorum gen. nov., sp. nov., a novel gammaproteobacterium from marine coastal sediments. . *Int J Syst Evol Microbiol*, **54**, 669-673.
198. Pinhassi J, P.M., Macián MC, Lekunberri I, González JM, Pedrós-Alió C, Arahal DR. (2007) Reinekea blandensis sp. nov., a marine, genome-sequenced gammaproteobacterium. . *Int J Syst Evol Microbiol*, **57**, 2370-2375.
199. Choi A, C.J. (2010) Reinekea aestuarii sp. nov., isolated from tidal flat sediment. *Int J Syst Evol Microbiol*, **60**, 2813-2817.
200. Giebel HA, K.D., Lemke A, Thole S, Gahl-Janssen R, Simon M, Brinkhoff T. (2011) Distribution of Roseobacter RCA and SAR11 lineages in the North Sea and characteristics of an abundant RCA isolate. . *ISME J*, **5**, 8-19.
201. Giovannoni SJ, V.K. (2012) Seasonality in ocean microbial communities. . *Science*, **335**, 671-676.
202. Hutchinson GE. (1961) The paradox of the plankton. . *Am Nat*, **95**, 137-145.
203. DeLong EF. (2009) The microbial ocean from genomes to biomes. *Nature*, **459**, 200-206.

204. Arrigo K. (2005) Marine microorganisms and global nutrient cycles. *Nature*, **437**, 349-355.
205. Ogura A, L.M., Shigenobu Y, Fujiwara A, Ikeo K, Nagai S. (2011) Effective gene collection from the metatranscriptome of marine microorganisms. *BMC Genomics*, **12**, S15.
206. Gilbert JA, M.F., Schriml L, Joint IR, Mühling M, Field D. (2010) Metagenomes and metatranscriptomes from the L4 long-term coastal monitoring station in the Western English Channel. *Stand Genomic Sci*, **3**, 183-103.
207. Li X, Q.L. (2005) Metagenomics-based drug discovery and marine microbial diversity. . *Trends Biotechnol*, **23**, 539-543.
208. DeLong EF, P.C., Mincer T, Rich V, Hallam SJ, Frigaard NU, Martinez A, Sullivan MB, Edwards R, Brito BR, Chisholm SW, Karl DW,. (2006) Community genomics among stratified microbial assemblages in the ocean's interior. *Science*, **311**, 496-503.
209. Williams TJ, L.E., Evans F, Demaere MZ, Lauro FM, Raftery MJ, Ducklow H, Grzymski JJ, Murray AE, Cavicchioli R. (2012) A metaproteomic assessment of winter and summer bacterioplankton from Antarctic Peninsula coastal surface waters. *ISME J*, **doi: 10.1038/ismej.2012.28.** .
210. Kan J, H.T., Ginter JM, Wang K, Chen F. (2005) Metaproteomic analysis of Chesapeake Bay microbial communities. *Saline Systems*, **1**.
211. Taverna DM, G.R. (2002) Why are proteins marginally stable? *Proteins*, **46**, 105-109.
212. Simon C, D.R. (2011) Metagenomic Analyses: Past and Future Trends. *Appl Environ Microbiol*, **77**, 1153-1161.
213. Klindworth A, P.E., Peplies J, Quast C, Horn M, Glöckner FO. (2012) Evaluation of general 16S ribosomal RNA gene PCR primers for classical and next generation sequencing based diversity studies. *Nucleic Acids Res*, **accepted**.
214. Chomczynski P, S.N. (1987) Single-step method of RNA isolation by acid guanidinium thiocyanate-phenol-chloroform extraction. *Anal Biochem*, **162**, 156-159.
215. Ning Z, C.A., Mullikin JC. (2001) SSAHA: a fast search method for large DNA databases. *Genome Res*, **11**, 1725-1729.
216. Kanehisa M. (2002) The KEGG database. *Novartis Found Symp*, **247**, 91-101.
217. Finn RD, M.J., Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer EL, Eddy SR, Bateman A. (2010) The Pfam protein families database. . *Nucleic Acids Res*, **38**, D211-222.
218. Giovannoni SJ, B.L., Cho JC, Stapels MD, Desiderio R, Vergin KL, Rappé MS, Laney S, Wilhelm LJ, Tripp HJ, Mathur EJ, Barofsky DF. (2005) Proteorhodopsin in the ubiquitous marine bacterium SAR11. *Nature*, **438**, 82-85.
219. Kemp PF, L.S., Laroche J. (1993) Estimating the growth rate of slowly growing marine bacteria from RNA content. *Appl Environ Microbiol*, **59**, 2594-2601.
220. Klappenbach JA, D.J., Schmidt TM. (2000) rRNA operon copy number reflects ecological strategies of bacteria. *Appl Environ Microbiol*, **66**, 1328-1333.
221. Noinaj N, G.M., Barnard TJ, Buchanan SK. (2010) TonB-dependent transporters: regulation, structure, and function. *Annu Rev Microbiol*, **64**, 43-60.
222. Tang K, J.N., Liu K, Zhang Y, Li S. (2012) Distribution and Functions of TonB-Dependent Transporters in Marine Bacteria and Environments: Implications for Dissolved Organic Matter Utilization. *PLoS One*, **7**, e41204.
223. Martens EC, K.N., Smith TJ, Gordon JI. (2009) Complex glycan catabolism by the human gut microbiota: the Bacteroidetes Sus-like paradigm. *J Biol Chem*, **284**, 24673-24677.
224. Yu K, Z.T. (2012) Metagenomic and Metatranscriptomic Analysis of Microbial Community Structure and Gene Expression of Activated Sludge. . *PLoS One*, **7**, e38183.

225. Janausch IG, Z.E., Tran QH, Kröger A, Unden G. (2002) C4-dicarboxylate carriers and sensors in bacteria. *Biochim Biophys Acta.*, **1553**, 39-56.
226. Thrash JC, C.J., Vergin KL, Morris RM, Giovannoni SJ. (2010) Genome Sequence of Lentisphaera araneosa HTCC2155T, the Type Species of the Order Lentisphaerales in the Phylum Lentisphaerae. *J Bacteriol*, **192**, 2938-2939.
227. Palmer KL, C.K., Manson JM, Heiman D, Shea T, Young S, Zeng Q, Gevers D, Feldgarden M, Birren B, Gilmore MS. (2010) High-quality draft genome sequences of 28 Enterococcus sp. isolates. *J Bacteriol*, **192**.
228. Henrissat B, B.A. (1996) Updating the sequence-based classification of glycosyl hydrolases. *316*, **Pt 2**.
229. Davies G, H.B. (1995) Structures and mechanisms of glycosyl hydrolases. *Structure*, **3**, 853-859.
230. Warren RA. (1996) Microbial hydrolysis of polysaccharides. *Annu Rev Microbiol*, **50**, 183-212.
231. Koropatkin NM, S.T. (2010) SusG: a unique cell-membrane-associated alpha-amylase from a prominent human gut symbiont targets complex starch molecules. *Structure*, **18**, 200-215.
232. Hong Cho K, S.A. (2001) Biochemical analysis of interactions between outer membrane proteins that contribute to starch utilization by Bacteroides thetaiotaomicron. . *J Bacteriol*, **183**, 7224-7230.
233. Wanner BL. (1993) Gene Regulation by Phosphate in Enteric Bacteria. . *J Cell Biochem*, **51**, 47-54.
234. Sayers EW, B.T., Benson DA, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Edgar R, Federhen S, Feolo M, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Madden TL, Maglott DR, Miller V, Mizrachi I, Ostell J, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Yaschenko E, Ye J. (2009) Database resources of the National Center for Biotechnology Information. *Nucleic Acids Res*, **37**, D5-15.
235. Gade D, G.J., Rabus R. (2005) Proteomic analysis of carbohydrate catabolism and regulation in the marine bacterium Rhodopirellula baltica. . *Proteomics*, **5**, 3672-3683.
236. Frank CS, L.P., Fuchs BM, Harder J. (2011) Ammonia and attachment of Rhodopirellula baltica. . *Arch Microbiol*, **193**, 365-372.
237. Müller I, K.A., Pape T, Sheldrick GM, Meyer-Klaucke W, Dierks T, Kertesz M, Usón I. (2004) Crystal structure of the alkylsulfatase AtsK: Insights into the catalytic mechanism of the Fe(II) alpha-ketoglutarate-dependent dioxygenase superfamily. . *Biochemistry*, **43**, 3075-3088.
238. Hagelueken G, A.T., Wiehlmann L, Widow U, Kolmar H, Tümmler B, Heinz DW, Schubert WD. (2006) The crystal structure of SdsA1, an alkylsulfatase from Pseudomonas aeruginosa, defines a third class of sulfatases. . *Proc Natl Acad Sci U S A*, **103**, 7631-7636.
239. Ghosh D. (2007) Human sulfatases: a structural perspective to catalysis. *Cell Mol Life Sci*, **64**, 2013-2022.
240. Dierks T, L.M., Schlotterhose P, Schmidt B, von Figura K. (1999) Sequence determinants directing conversion of cysteine to formylglycine in eukaryotic sulfatases. . *EMBO J*, **18**, 2084-2091.
241. Berteau O, G.A., Benjdia A, Rabot S. (2006) A new type of bacterial sulfatase reveals a novel maturation pathway in prokaryotes. *J Biol Chem*, **281**, 22464-22470.
242. Benjdia A, D.G., Rabot S, Berteau O. (2007) First evidences for a third sulfatase maturation system in prokaryotes from E. coli aslB and ydeM deletion mutants. . *FEBS Lett*, **581**, 1009-1014.

243. Carlson BL, B.E., Skordalakes E, King DS, Breidenbach MA, Gilmore SA, Berger JM, Bertozzi CR. (2008) Function and structure of a prokaryotic formylglycine-generating enzyme. . *J Biol Chem*, **283**, 20117-20125.
244. Hanson SR, B.M., Wong CH. (2004) Sulfatases: Structure, mechanism, biological activity, inhibition, and synthetic utility. . *Angew Chem Int Ed Engl*, **43**, 5736-5763.
245. Hieu CX, V.B., Albrecht D, Becher D, Lombardot T, Glöckner FO, Amann R, Hecker M, Schweder T. (2008) Detailed proteome analysis of growing cells of the planctomycete Rhodopirellula baltica SH1T. *Proteomics*, **8**, 1608-1623.
246. Woebken D, T.H., Wecker P, Dumitriu A, Kostadinov I, Delong EF, Amann R, Glöckner FO. (2007) From the Namibian and Oregon coast upwelling systems and their cross-comparison with planctomycete genomes. . *ISME J*, **1**, 419-435.
247. Lage OM, B.J. (2011) Planctomycetes diversity associated with macroalgae. *Environ Microbiol*, **78**, 366-375.
248. Bengtsson MM, S.K., Øvreås L. (2010) Seasonal dynamics of bacterial biofilms on the kelp Laminaria hyperborea. *Aquat Microbiol Ecol*, **60**, 71-83.
249. Lahaye M, R.A. (2007) Structure and functional properties of ulvan, a polysaccharide from green seaweeds. *Biomacromolecules*, **8**, 1765-1774.
250. Usov AI, B.M. (2009) Fucoidans - sulfated polysaccharides of brown alga. *Russ Chem Rev*, **78**.
251. Michel G, N.-C.P., Barbeyron T, Czjzek M, Helbert W. (2006) Bioconversion of red seaweed galactans: a focus on bacterial agarases and carrageenases. . *Appl Microbiol Biotechnol*, **71**, 23-33.
252. Wecker P, K.C., Schüler M, Dabin J, Michel G, Glöckner FO. (2010) Life cycle analysis of the model organism Rhodopirellula baltica SH 1T by transcriptome studies. *Microb Biotechnol*, **3**, 583-594.
253. Winkelmann N, H.J. (2009) An improved isolation method for attached-living Planctomycetes of the genus Rhodopirellula. . *J Microbiol Methods*, **77**, 276-284.
254. Winkelmann N, J.U., Meyer C, Serrano W, Rachel R, Rosselló-Mora R, Harder J. (2010) Determination of the diversity of Rhodopirellula isolates from European seas by multilocus sequence analysis. . *Appl Environ Microbiol*, **76**, 776-785.
255. Richter M, R.-M.R. (2009) Shifting the genomic gold standard for the prokaryotic species definition. *Proc Natl Acad Sci U S A*, **106**, 19126-19131.
256. Frank CS. (2011) Polyphasische Taxonomue, Kerngenom und Lebenszyklus von Rhodopirellula-Stämmen. *Ph.D. thesis - University of Bremen, Germany*.
257. Li L, S.C.J., Roos DS. (2003) OrthoMCL: identification of ortholog groups for eukaryotic genomes. *Genome Res*, **13**, 2178-2189.
258. Kumar AS, M.K., Jha B. (2007) Bacterial exopolysaccharides - a perception. *J Basic Microbiol*, **47**, 103-117.
259. Waterhouse AM, P.J., Martin DM, Clamp M, Barton GJ. (2009) Jalview Version 2-a multiple sequence alignment editor and analysis workbench. . *Bioinformatics*, **25**, 1189-1191.
260. Stamatakis A. (2006) RAxML-VI-HPC: Maximum Likelihood-based phylogenetic analyses with thousands of taxa and mixed models. . *Bioinformatics*, **22**, 2688-2690.
261. Miller MA, P.W., Schwartz T. . (2010) Creating the CIPRES Science Gateway for inference of large phylogenetic trees. . *In: Proceedings of the Gateway Computing Environments Workshop (GCE)*, **14. Nov**, LA pp 1 - 8. .
262. Han MV, Z.C. (2009) phyloXML: XML for evolutionary biology and comparative genomics. *BMC Bioinformatics*, **10**, 356.
263. Crooks GE, H.G., Chandonia JM, Brenner SE. (2004) WebLogo: A sequence logo generator. . *Genome Res*, **14**, 1188-1190.

264. Battke F, S.S., Nieselt K. (2010) Mayday - integrative analytics for expression data. . *BMC Genomics*, **11**.
265. Katoh K, M.K., Kuma K, Miyata T. (2002) MAFFT: a novel method for rapid multiple sequence alignment based on fast Fourier transform. *Nucleic Acids Res*, **30**, 3059-3066.
266. Bendtsen JD, N.H., von Heijne G, Brunak S. (2004) Improved prediction of signal peptides: SignalP 3.0. *J Mol Biol*, **340**, 783-795.
267. Krogh A, L.B., von Heijne G, Sonnhammer EL. (2001) Predicting transmembrane protein topology with a Hidden Markov Model: Application to complete genomes. *J Mol Biol*, **305**, 567-580.
268. Zierer MS, M.P. (2000) A wide diversity of sulfated polysaccharides are synthesized by different species of marine sponges. . *Carbohydr Res*, **328**, 209-216.
269. Ziervogel K, A.C. (2008) Polysaccharide hydrolysis in aggregates and free enzyme activity in aggregate-free seawater from the north-eastern Gulf of Mexico. . *Environ Microbiol*, **10**, 289-299.
270. Bilan MI, G.A., Shashkov AS, Nifantiev NE, Usov AI. (2006) Structure of a fucoidan from the brown seaweed Fucus serratus. . *Carbohydr Res*, **341**, 238-245.
271. Li B, L.F., Wei X, Zhao R. (2008) Fucoidan: Structure and bioactivity. . *Molecules*, **13**, 1671-1695.
272. Jogler C, G.F., Kolter R. (2011) Characterization of Planctomyces limnophilus and development of genetic tools for its manipulation establish it as a model species for the phylum Planctomycetes. *Appl Environ Microbiol*, **77**, 5826-5829.
273. Deutscher J, F.C., Postma PW. (2006) How phosphotransferase system-related protein phosphorylation regulates carbohydrate metabolism in bacteria. . *Microbiol Mol Biol Rev*, **70**, 939-1031.
274. Siebold C, F.K., Beutler R, Erni B. (2001) Carbohydrate transporters of the bacterial phosphoenolpyruvate: sugar phosphotransferase system (PTS). . *FEBS Lett*, **504**, 104-111.
275. Poli A, A.G., Nicolaus B. (2010) Bacterial exopolysaccharides from extreme habitats: production, characterization and biological activities. . *Mar Drugs*, **8**, 1779-1802.
276. Loy A, M.F., Wagner M, Horn M. . (2007) probeBase - an online resource for rRNA-targeted oligonucleotide probes: new features 2007. *Nucleic Acids Res*, **35**, D800-D804.
277. Moraru C, L.P., Fuchs BM, Kuypers MM, Amann R. (2010) GeneFISH--an in situ technique for linking gene presence and cell identity in environmental microorganisms. *Environ Microbiol*, **12**, 3057-3073.
278. Yilmaz P, G.J., Knight R, Amaral-Zettler L, Karsch-Mizrachi I, Cochrane G, Nakamura Y, Sansone SA, Glöckner FO, Field D,. (2011) The genomic standards consortium: bringing standards to life for microbial ecology. . *ISME J*, **5**, 1565-1567.
279. Kennedy J, O.L.N., Kiran GS, Morrissey JP, O'Gara F, Selvin J, Dobson AD. (2011) Functional metagenomic strategies for the discovery of novel enzymes and biosurfactants with biotechnological applications from marine ecosystems. *Journal of Applied Microbiology*, **111**, 1365-2672.