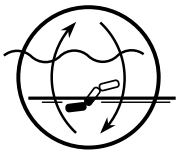


Max Planck Institute
for Marine Microbiology

JACOBS
UNIVERSITY

School of Engineering and Science

Marine Metagenomics

From high-throughput data to ecogenomic interpretation

by

Ivaylo Kostadinov, M.Sc.

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

DOCTOR OF PHILOSOPHY

in Bioinformatics

Approved Thesis Committee:

Prof. Dr. Frank Oliver Glöckner (chair)
Max Planck Institute for Marine Microbiology
Jacobs University

Prof. Dr. Matthias Ullrich
Jacobs University

PD Dr. Bernhard Fuchs
Max Planck Institute for Marine Microbiology

Date of Defense: 6 May 2011

When we try to pick out anything by itself, we find it hitched to everything else in the universe.

John Muir (1838 - 1914)

Thesis Abstract

The field of marine ecological genomics is evolving at an unprecedented pace. The increasingly cheaper and faster sequencing technologies present new possibilities but also new challenges. The sheer amount of sequencing data requires elegant solutions for storing, querying and exchanging it. Additionally, sequences alone are not enough to understand the complex interactions between microorganisms and the marine environment. Comprehensive environmental and contextual metadata are needed to put the sequences into context. However, the integration of sequence and metadata is not a trivial task.

The aim of this thesis was to enhance the field of marine ecological genomics by fulfilling two tasks. Firstly, the technology necessary for data integration, visualization and analysis was set up. Secondly, this technological platform was used to support three ecological analyses of marine microbes. The first task was addressed by improving the existing portal for marine ecological genomics (http://www.megx.net). The main goal of megx.net is to integrate sequences and their metadata based on geographic location. The focus lies on environmental parameters such as temperature, salinity and nutrients. As a result of this thesis, megx.net profits from a new data model, a new web interface, improved visualization and analysis tools. For the second task, the integrated resources offered by megx.net were used. The effect of environment stability on the transcription factors content of microbial communities was quantified using interpolated environmental data. In a second study, interpolations were used to give environmental context to new hypothesis for domains of unknown function in the marine metagenomes. Last but not least, a study of the community structures of high- and low-DNA content marine microbes was complemented with integrated metadata.

The extended technological platform megx.net was successfully used to gain new insights into the ecology of marine microbes.

Contents

List of Figures

CHAPTER 1

INTRODUCTION

1.1 Marine Microbes

1.1.1 Ecology and impact

In the history of biological sciences, the importance of prokaryotes has only recently been recognized. Antonie van Leeuwenhoek, the founder of microbiology, discovered 'miniscules' in the 17th century. It took his successors about 200 years to realize how massive the distribution and the impact of these microorganisms really is. Following the work of Robert Koch, pathogenic microbes have been under close scrutiny for over 100 years now. Once bacteria-induced diseases were understood, new challenges arose. Some of them, like anthropogenic impact on global processes (e.g. food chains, climate), are amongst the most discussed topics of our time. In order to address these topics, we need to understand the complex nature of microbes and how they interact with their environment.

Largely underestimated at first, the total number of prokaryotes on Earth is now estimated to be around 4–6 $\times$ 10^{30}. Their carbon content is close to that of all other living organisms combined. Despite their miniature size, prokaryotes comprise the largest pool of organic nitrogen and phosphorus [Whitman et al., 1998]. The largest bacterium known to date, *Thiomargarita namibiensis*, is about 750 μm in diameter [Schulz et al., 1999]. The tiniest *Archaea*, *Thermodiscus*, reach only 0.2 μm in diameter. Differences in volume cover a range of 10 orders of magnitude [Schulz and Jorgensen, 2001]. Microbes can be found almost everywhere and the human body is no exception. It harbors somewhere between 10 and 100 times more bacterial cells than the eukaryotic cells it is made of (estimated 10^{13}) [Savage, 1977, Berg, 1996]. Microorganisms inhabit even the harshest environments: hy-

persaline, hyperthermal, and highly acidic [Kivistö and Karp, 2010, Huber et al., 2000, Tyson et al., 2004]. The diversity of niches microorganisms occupy can probably only be matched by their metabolic diversity. Microbes are capable of using different energy sources (e.g. light, inorganic chemicals), under various conditions (e.g. aerobic, anaerobic) to utilize both organic and inorganic carbon sources [Madigan et al., 2003]. They not only produce the compounds they need but can also make them available for other micro- and macro-organisms. A prominent example is nitrogen fixation, where atmospheric nitrogen (N_2) is converted to ammonia (NH_3) which is essential for protein synthesis. This process is only carried out by prokaryotes. Some of them are free-living but many have been 'adopted' as symbionts by eukaryotic organisms [Fiore et al., 2010, Francis et al., 2007].

About 10 years ago, the world ocean was recognized to be one of the biggest reservoirs of microorganisms. It harbors a total of 10^{29} microbial cells or around 10^5 cells per milliliter of seawater [Whitman et al., 1998]. Many of the important processes take place in the upper 200 meters of ocean water where approximately 3×10^{27} autotrophic microbes live. Autotrophic organisms are capable of producing organic matter from inorganic carbon. This process is also known as primary production. Microbial-driven primary production is essential for all life on earth: marine prokaryotes have high nutrient turnover rates and fix the same amount of CO_2 as terrestrial plants [Field, 1998, Woodward, 2007]. The global ocean is the major reservoir of inorganic carbon, holding 50 times more of it than the atmosphere. Atmospheric CO_2 concentrations depend on the equilibrium transport to and from the ocean. Both biotic and abiotic factors influence the solubility of inorganic carbon in marine water. The abiotic factors are temperature, alkalinity and salinity of the surface waters [Raven J. A. and Falkowski, 1999]. The biotic factors include the incorporation of carbon into biomass and is more complicated. For example, carbon is constantly being removed form the surface to the deep ocean where it is trapped on geological scales (millions of years). This phenomenon, known as the biological pump (BP), is based on sinking particles. Although the BP keeps the levels of atmospheric CO_2 down to half of what they would be without it, it is not a perfect process. Cell lysis by grazers and especially by viral infection converts particulate organic matter (POC) back

to dissolved. It is then available for re-use by living organisms or gets transported to the atmosphere. The role of viruses in releasing the trapped POC is known as the viral shunt [Suttle, 2005, Suttle, 2007]. Since the industrial revolution, the human input of CO_2 to the atmosphere has been substantial and has given grounds for much debate and concern. Therefore, the role of the ocean as a natural sink for CO_2 is of great interest. Iron fertilization experiments attempt to boost the uptake of atmospheric CO_2 by causing artificial phytoplankton blooms in iron-limited areas of the ocean. So far the results are controversial about the efficacy and large-scale applicability of such [Raven J. A. and Falkowski, 1999, Smetacek and Naqvi, 2008]. This comes to show that we need the thorough understanding of marine microorganisms, before we can harvest their abilities.
Marine microorganisms have a far-reaching impact on our lives. Microbial ecologists struggle to better understand this impact, learn to reliably predict it and maybe, in time, to steer it. The complex nature of microorganisms, their metabolic abilities, community structures and interactions with the environment are far from trivial to determine. It requires an interdisciplinary approach covering fields from biology to informatics.

1.1.2 Towards culture-independent interrogation of microbial communities

Unlocking the secrets of marine microbes requires the application of a vast array of techniques and approaches. Classical microbiology deals with cultivation and characterization of organisms. It is able to address but a tiny portion of the microorganisms we know exist. The vast majority of them (95-99%) remain 'unseen' in pure culture [Amann et al., 1995]. It is very likely that specific conditions from their habitats that cannot be mirrored in the laboratory. Symbiotic dependencies between members of the communities are perfect examples of such conditions. Marine surface water communities were shown to be dominated by bacteria which are adapted to oligothrophic conditions and are less amenable to cultivation [Lauro et al., 2009]. Instead of being discouraging, these facts prompt the use of different techniques to learn the more about the unculturable majority in order to surpass the cultiva-

tion barrier. For example, using molecular biology and genomics to establish their community structure and metabolic capabilities. Well established molecular techniques like Fluorescence In-Situ Hybridization [DeLong et al., 1989, Amann et al., 1990] have proven indispensable in characterizing the community structure of planktonic prokaryotes [Grossart et al., 2005, Schattenhofer et al., 2009, Simonato et al., 2010]. Nevertheless, many molecular techniques are time- and effort-consuming, making large-scale comparative studies hard to carry out. Their results depend largely on the protocols used, making comparison between similar studies difficult. The basic culture-independent approach to microbial diversity was developed as far back as the 1980s [Pace et al., 1985]. PCR technology eliminated the culturing bottleneck for microbial diversity studies. The uncultured majority of microorganisms were shown to be highly diverse and sometimes only distantly related to the culturable few [Rappé and Giovannoni, 2003]. The culture-independent methods for accessing microbial diversity, despite offering many new insights and advantages, do not invalidate culturing efforts. On the contrary, the two complement each other. Members of the SAR11 clade represent around a third of the prokaryotic cells in the ocean surface waters. This discovery, based on 16S rRNA signatures, prompted even more culturing efforts, in order to more properly describe these key marine organisms [Handelsman, 2004].
Culture independent methods for diversity analysis are a major step towards better understanding of marine microbial communities. Once the major members of the community are identified, the next questions is what functional repertoire they have. To give an answer, community genomics, or metagenomics, is used.

1.2 Metagenomics

1.2.1 From single genes to community genomics

Genomics is the study of sequenced hereditary material (DNA). Its foundations were laid in the late 1860s with the discovery of DNA [Dahm, 2007]. However, it was not until the early 1970s before the first RNA bases were sequenced [JOU et al., 1972], and another 20 years

until the launch of the Human Genome Project in 1990. It took about 10 years to produce a draft of the human genome. The technologies and experience gathered along the way benefited the whole field of genomics. Meanwhile in 1995, the first complete genome was sequenced. The organism was *Haemophilus influenzae*, a pathogenic bacterium which infects humans [Fleischmann et al., 1995]. Only 10 years later genomics had evolved quickly from studying single genes to comparing several hundred genomes. However, the culturing bias influences greatly which complete genomes are available [Pace, 1997]. Easily culturable and highly abundant species are far better represented. The Microbial Earth Project[1] aims to obtain a draft genome from all available type strains. Its pilot program, the Genomic Encyclopedia of *Bacteria* and *Archaea* (GEBA), showed promising results on the way to 'fill the gaps' in the phylogenetic tree. [Wu et al., 2009]. An alternative approach to circumvent the culturing bias is to sequence and analyze whole microbial communities, a technique know as metagenomics.
Metagenomics is a term coined by Handelsman and coworkers in 1998 [Handelsman et al., 1998]. It describes the sequencing and analysis of whole microbial communities from environmental samples. Community genomics, environmental genomics and population genomics are often used as synonyms. Direct cloning of DNA from environmental samples was proposed in the late 80s [Pace et al., 1985]. Metagenomics involves DNA isolation from an environmental sample, cloning of the DNA into a vector, and transforming the clones into a host bacterium. Depending on the scientific questions several approaches can be taken from here on. The transformed clones can be screened for phylogenetic markers (e.g. 16S rRNA) of a target organism or clade. Once these are found, the respective clones can be completely sequenced to reveal the functional potential of the target clade. Reversely, one could screen for a functional gene of interest first and try to identify the responsible organisms as a next step. Since the sequencing revolution, either random sequencing of clones or high throughput sequencing of complete DNA libraries offer unprecedented insights into community structure and functional potential on different scales [Riesenfeld et al., 2004, Handelsman, 2004]. Moreover, genome reconstruction

[1] `http://genome.jgi-psf.org/programs/bacteria-archaea/MEP/index.jsf`

from metagenomic samples can narrow down the analysis to specific clades or single organisms in a culture-independent way [Venter et al., 2004, Meyerdierks et al., 2010]. Such analysis might reveal key parameters necessary for the successful isolation of yet unculturable organisms. Examples of the unexpected findings in metagenomic data are ample. In a cornerstone work, DeLong and coworkers reported the discovery of an archaeal 16S rRNA gene in a metagenomic library constructed from seawater [Stein et al., 1996]. Bacterial rhodopsins were found in a uncultured γ-Proteobacterium [Béjà et al., 2000]. It proved marine autotrophs possess a light-driven proton pump based on other pigments than chlorophyll. Further metagenomic studies revealed high diversity of bacterial proteorhodopsins [Venter et al., 2004]. As any other technique, metagenomics presents some challenges. Among them are the library size, the detection of phylogenetic anchors and the lack of functional confirmation [Riesenfeld et al., 2004, Warnecke and Hugenholtz, 2007]. The size of a metagenomic library determines the coverage of the genetic material and the likelihood to also cover the rare species in a sample. A metagenomic library from 1 ml of seawater should have a size of approximately 500 Gbp to properly depict the species richness, including the rare members of the community [Riesenfeld et al., 2004]. Libraries from such sizes are costly to prepare and produce enormous amounts of data. Linking the metabolic potential of a community to the identity of its members often relies on finding and correctly identifying a phylogenetic marker like the 16S rRNA gene. However, the number of bacterial rRNA operons per cell varies form 1 to 15 and is influenced by growth rate [Klappenbach et al., 2000]. This means that slow growing, difficult to culture bacteria will be under-represented in 16S libraries. Ways to circumvent such problems include increasing the sequence coverage, followed by some level of assembly, and using several phylogenetic markers simultaneously. Assembling metagenomic fragments is not a trivial task, especially for short reads (75-100bp) which are common for current sequencing technologies (e.g. Illumina). The issue of identifying community members is being addressed in several ways. Complex communities can be separated into subsets by applying molecular techniques such as fluorescence-activated cell sorting [Warnecke and Hugenholtz, 2007]. Such studies have already broadened our understanding of

marine microbial communities [Kalyuzhnaya et al., 2008, Woyke et al., 2009, Tripp et al., 2010, Schattenhofer et al., 2011]. Multiple displacement amplification (MDA) is a method to amplify DNA starting with very little template. It enables single cell genomics and allows genomic access to the rare members of microbial communities [Binga et al., 2008]. Next-generation sequencing technologies promise faster, cheaper, more accurate sequences with longer read length [Metzker, 2009]. This would allow better assembly and the application of established *in silico* methods which are designed for longer sequences. In the mean time, novel *in silico* applications make use of Self-Organizing Maps for binning and phylogenetic classification of short sequencing reads [Martin et al., 2008, Weber et al., 2011]. Last but not least, metagenomic data can only describe the functional potential of microbial communities. If and when this potential is realized is a question to be answered by investigating the community's transcriptome and proteome.

Metagenomics has revolutionized marine microbial ecology. It offers unprecedented insights into the gene pool and diversity of microbial communities, which are inaccessible through culture-dependent genomics. Despite its limitations, metagenomics is one of the most consequential techniques of our time.

1.2.2 Ecological multi-omics

Some of the basic questions in microbial ecology have been around for a long time: What is the structure of microbial communities in the environment? What functions can these communities perform? How do organisms interact with one another and with their environment? Ecological genomics tries to answer these questions by interrogating community genomics data. However, to fully understand the ecology of microorganisms, additional methods must be used. The suffix '-omics' is often applied to different fields of molecular biology to give the meaning of "studying all the parts together". Besides genomics, transcriptomics and proteomics are the most relevant techniques for studying marine microbial ecology.

Transcriptomics studies all the transcripts present in the cell at any time. These include messenger RNAs (mRNAs), ribosomal RNAs (rRNAs),

transfer RNAs (tRNAs) and small nuclear RNAs (snRNAs). Transcriptomics focuses primarily on mRNAs and their relation to certain conditions that influence their expression: for example which genes are expressed during the life cycle or which genes are up/down regulated during feast and famine. Transcriptomics is a functional analysis because it studies the expressed functions in a cell, as compared to the potential encoded in its genome (the field of genomics). The transcriptome depends on the environmental conditions, physiological state, developmental state and many other factors. Transcriptomics gives a more concrete picture which genes are "active" at a certain time [van Straalen and Roelofs, 2006]. The time frame is a limiting factor: an interesting or important function might not be active at the time the sample was taken and will not be detected by transcriptomics. Another major limitation is the short lifetime of mRNA. It makes RNA extraction a difficult task where time is of the essence.
Proteomics is a study which targets the entire protein content of organisms. It was made possible by advances in the mass spectrometry analysis. It allows to fingerprint the proteins according to their mass. The proteome of an organism often differs significantly from its transcriptome. The reason is translational control which can be dictated by physiological adaptation. Post-translational modification of proteins is another mechanism contributing to this difference. The proteome and the genome are connected through a complex feedback network. Some proteins function as DNA binding transcription factors that influence the activation or repression of genes. Others are involved directly in transcription or translation. Still others are structural components of chromosomes. If the aim of molecular biology is to study the cell in full, then combining genomics with transcriptomics and proteomics is the minimum effort required [van Straalen and Roelofs, 2006].
Both transcriptomics and proteomics have their extensions to study whole microbial communities. Metatranscriptomics has contributed to the better characterization of functional and taxonomic diversity of marine microbes, their role in the carbon cycle and the discovery of novel gene categories [Gilbert et al., 2010, Frias-Lopez et al., 2008, McCarren et al., 2010]. Metaproteomics has been used to study the functional response of marine prokaryotes to different nutrient conditions [Morris et al., 2010, Sowell et al., 2011]. Further 'omics' approaches, focusing

on protein interactions and metabolites, are also used to study microbial systems [Zhang et al., 2010].
A thorough understanding of marine microbial communities requires an integrative approach where results form metagenomic, metatranscriptomic and metaproteomic studies are combined. In order to interpret these results in the ecological context we are interested in, a detailed, precise and comparable description of the environment is essential.

1.3 Metadata and Metaanalysis

1.3.1 Ecological genomics as an integrative science

Science, research, technology, innovation. These terms describe different aspects of the same basic drive: struggle for knowledge. What comes out of a sequencing machine is nothing more than raw data. It has to be processed to arrive at information, which can be modeled into knowledge. Then goals can be added to arrive at wisdom. Finally, values are included to create a vision. It is estimated that along this way, less and less cognitive capacities are invested[2]. Applying this model to ecological genomics would make data collection and initial handling the most intellectually demanding. Indeed, any piece of data is only as good as the information that can be extracted from it. Consequently, in order to maximize the knowledge produced from ecogenomic data, careful planing of its storage, querying and exchange is needed. Advanced technological platforms for data integration can provide long term solutions to this task. This will allow us to concentrate our cognitive capacities on the analysis, visualization and interpretation of the data. Shifting the focus of our intellectual efforts is probably the key to fully understanding microbial communities in the ocean. But sequences alone are not enough and neither are the results of other multi-omics approaches. Omics results should be examined in broad ecological perspective, which is only possible after comprehensive integration with environmental and contextual metadata.

[2] http://www.cognitivecybernetics.com/PrimerFoU.html

1.3.2 Meta- approach to marine microbial ecology

Metadata is best defined as data about data. The word **meta** is of Greek origin ($\mu\epsilon\tau\alpha$) and one of its many meanings is "adjacent". Examples of metadata in genomics describe the sampling (e.g. date, time, location, methods) and environmental conditions at the time. Metadata is crucial to the interpretation of sequences in ecological perspective. Metadata is usually collected according to the research plan of the scientist collecting the samples. Later on, different questions might evolve that require metadata that has not been collected. Our own experience in megx.net project shows that only 5.3% of the completely sequenced genomes have are georeferenced (i.e. the GPS coordinates of the sampling location are known). Metadata is a prerequisite for interpreting the ever growing metagenomic datasets. The Genomic Standards Consortium (GSC) is developing a set of specifications for the minimum information required to describe genomes, metagenomes and marker genes [Field et al., 2008, Yilmaz et al., 2011b]. Namely, these are the minimum information about a genome sequence (MIGS), its extension to the minimum information about a metagenome sequence (MIMS) and the minimum information about a marker gene sequence (MIMARKS). The standards are implemented in the form of checklists which are based on discussions between a broad range of scientists. Some data are considered mandatory and some are recommended. At the moment no public sequence resource enforces the checklist, making GSC compliance a matter of personal choice. The efforts of the GSC are being rewarded already. The International Nucleotide Sequence Databases Consortium (INSDC) has begun to incorporate the MIGS/MIMS/MIMARKS checklist as an additional structured text field in their submission forms. The first tools for consistent contextual data acquisition and submission are also available [Hankeln et al., 2010]. The whole process of integrating new standards into the big public resources requires time and effort. In the meanwhile, samples are still being taken, but scientists are advised to have the checklist in mind. Currently, numerous institutions comply to the checklist but none enforce it. The best level at which such quality control is applied is a matter of discussion. Some believe the public databases should be responsible. However, the submitting party has the final word. On the

other hand, metadata is most important for the interpretation of the sequences which is the basis of subsequent scientific publication. Therefore, it might be better to let scientific journals require the submission of MIGS/MIMS/MIMARKS compliant data to a public database before publication. This scheme has the additional advantage of putting pressure at the right spot: scientific output is still largely measured by publishing units. Standardized contextual data needs to be uniform not only in what is reported but how. A GSC project, Environment Ontology[3] (EnvO), provides a set of standardized terms to describe the environment where samples were taken from [Hirschman et al., 2008]. The advantage of using ontologies is the increased comparability between datasets annotated with the same set of terms [Yilmaz et al., 2011a].
Contextual metadata is of key importance for interpreting genomic sequences. It is often very dispersed and heterogeneous. Our own efforts to integrate sequence data from some of the most widely used public resources show that this is no easy task. Inconsistencies, mismatching data identifiers, different update cycles of resources are among the most common impediments for smooth data integration. Even with the appropriate standards in place, integration platforms have to be robust and flexible at the same time.

1.4 Bioinformatic Challenges in Ecological Genomics

Second generation sequence technologies (e.g. pyrosequencing) drastically changed genomics. They are immensely cheaper and faster than their predecessor, the Sanger sequencing technique, but produce significantly shorter reads [Hugenholtz and Tyson, 2008,Shendure and Ji, 2008]. The short read length is often seen as a disadvantage in ecological genomics. However, even 100bp sequences can sometimes suffice for microbial community analysis [Liu et al., 2007]. Different sequencing technologies can be combined for better results. A technology with comparatively longer read length (454 pyrosequencing) can be used to produce reference sequences with low coverage. High-throughput,

[3] http://environmentontology.org/

low-read length technology (e.g. Illumina) can be then used to achieve high coverage at low cost, by matching the short reads to the long reference ones. Such an approach has been successfully applied to assemble bacterial draft genomes [Croucher, 2009]. Sequence data from current technologies offers unprecedented possibilities to study microbial communities. However, its handling requires appropriate bioinformatic tools [Shendure and Ji, 2008, Metzker, 2009]. The main challenges posed by the huge amounts of second-generation sequence data can be roughly divided into (1) storage, integration, and exchange and (2) analysis, visualization, and interpretation. Bioinformatics applies computer science concepts to solve biological questions and helps tackle at least the first two of these.

1.4.1 Storage, integration and exchange

In the last years, newly emerged resources and techniques facilitate the use of high-throughput metagenomic data for exploring the ecology of marine microbes. The Community Cyberinfrastructure for Advanced Microbial Ecology Research and Analysis, in short CAMERA, offers a data repository and bioinformatic tools for microbial metagenomic data [Seshadri et al., 2007]. The Integrated Microbial Genomes (IMG) system and its extension for metagenomes (IMG/M) combines metagenomic data with isolate microbial genomes [Markowitz et al., 2008]. Next to the International Nucleotide Sequence Databases Collaboration (INSDC)[4] they are the main repositories for microbial metagenomic data. Data integration is of key importance for ecological genomics. The Marine Ecological Genomics portal (www.megx.net) focuses in the enrichment of isolate and metagenomic sequence data with metadata, especially environmental parameters [Kottmann et al., 2010]. Data integration in genomics is a task that is complicated by the heterogeneity of the data sources, general lack of standards for data description and exchange [Goble and Stevens, 2008]. Data integration requires curated data, but the way it is integrated should be curated as well [Goble et al., 2008]. Providing a unified view of data from heterogeneous sources can be realized in two ways. Either all data is collected in a centralized resource and updated regularly, or the original data providers agree on

[4] http://www.insdc.org/

a standardized exchange formats [Zhang et al., 2009]. The latter is highly dependent on Internet technologies such as Web Services (WS) and exchange formats. An extensible exchange format for MIGS/MIMS compliant genomic data is already in place [Kottmann et al., 2008]. Internet communication, offers the possibility to harvest the expertise of a wide user community. The so called 'Wiki' solutions are often used for community annotation, although their design is not entirely appropriate for the task [Arita, 2009].

1.4.2 Analysis, visualization and interpretation

All resources mentioned above offer a selection of tools to access, visualize and analyze the data. The metagenomics RAST server is specialized resource for automatic functional and phylogenetic annotation of metagenomic datasets [Meyer et al., 2008]. METAREP is a web 2.0 application for comparative metagenomics [Goll et al., 2010]. Diversity analysis is addressed by tools like ESPRIT [Sun et al., 2009] and the SILVA resource for quality-controlled rRNA [Pruesse et al., 2007]. These are just some prominent examples of a myriad of analysis tools for high-throughput microbial ecology.

Processing of metagenomic data is becoming so computationally intensive that the analysis is getting more expensive than the sequencing [Editorial, 2009]. The GenBank database estimates the doubling time of sequence data is around 18 months[5](Figure 1.1). Current and upcoming large sequencing projects will significantly shorten this period. Moreover, advanced sequencing techniques will produce more data faster (Figure 1.2) [Metzker, 2009]. Moore's law states that the doubling time of CPU transistors is approximately 24 months[6]. However, computational power does not rely only on CPU but also on other hardware components which do not improve with the same speed. In other words, sequence data generation is steadily out-competing the available computational resources and the trend seems to be irreversible. Solutions to this problem lie in the development of novel algorithms and the use of high-performance computing platforms. Homology searches have recently been accelerated up to 10^4 fold [Eddy,

[5] ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt
[6] http://en.wikipedia.org/wiki/Moore%27s_law

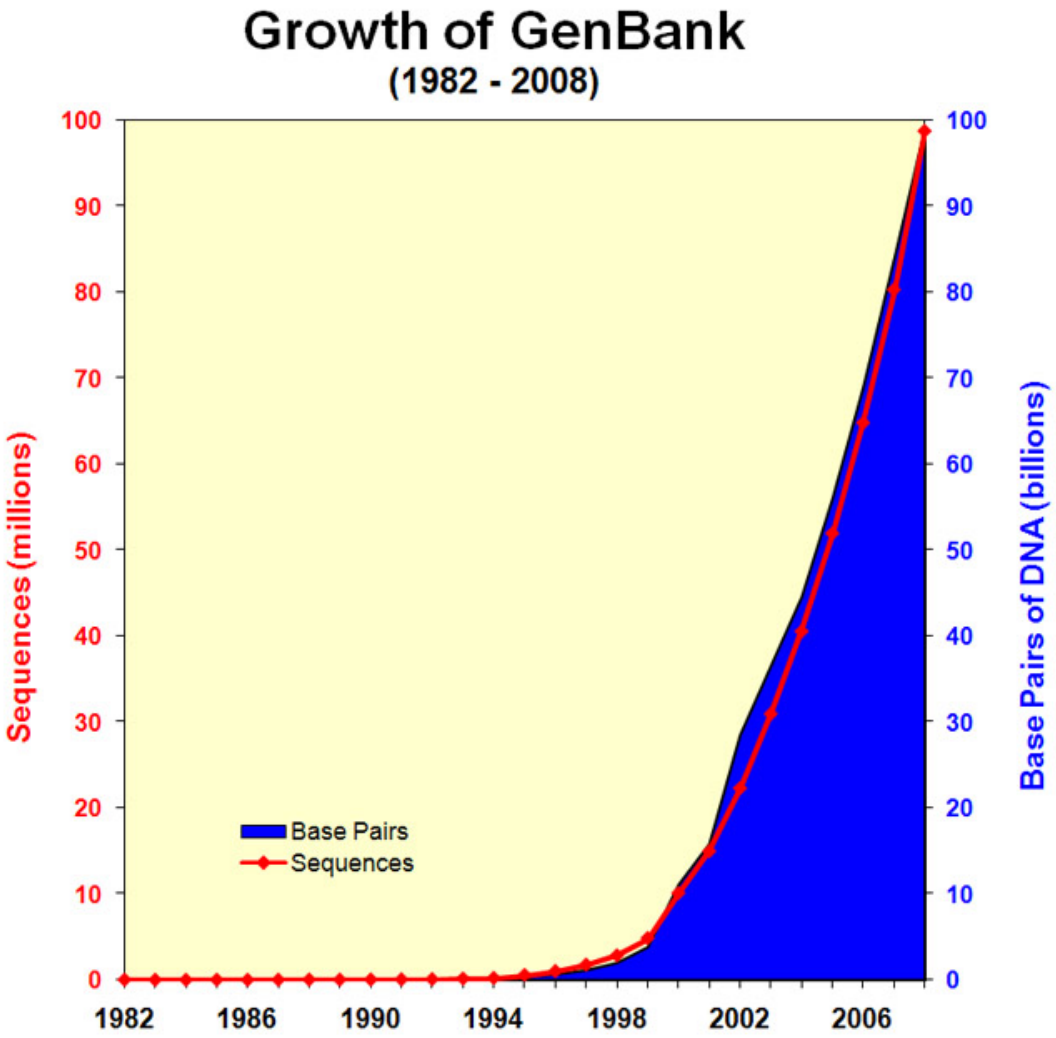


Figure 1.1: Increase of nucleotide sequence data in GenBank. Image from `http://www.ncbi.nlm.nih.gov/genbank/genbankstats.html`

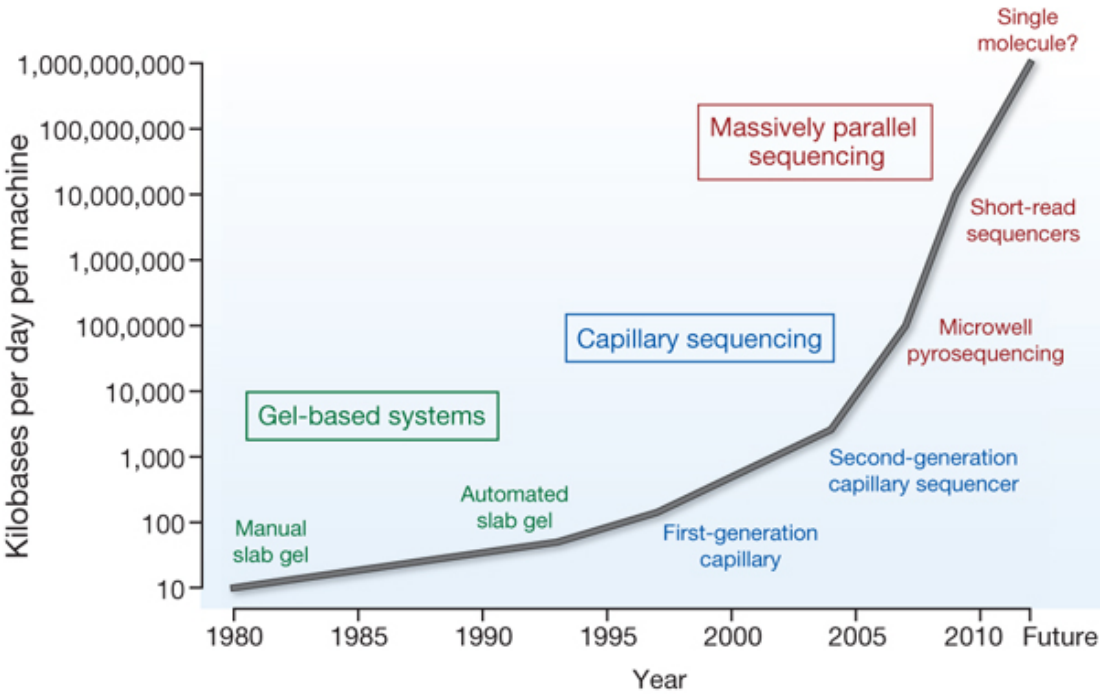


Figure 1.2: The capacity of sequencing technologies increases exponentially. Image from [Stratton et al., 2009]

2009, Meinicke, 2009]. To efficiently increase the speed and lower the cost of computation, such algorithms can be implemented directly in the hardware of graphic cards [Manavski and Valle, 2008]. Cloud computing can relax the computational bottleneck but cannot eliminate it completely. The first case studies with metagenomic data are already available [Schadt et al., 2010, Stein, 2010, Wilkening et al., 2009]. Last but not least, analysis of large metagenomic dataset is no longer possible without appropriate statistical techniques. Statistics is essential for removing biases and making meaningful ecological interpretations [Schloss and Handelsman, 2008, Beszteri et al., 2010, Parks and Beiko, 2010].

The low cost and high throughput of current and upcoming sequencing technologies has transformed ecological genomics into a highly data-intensive science. Sequence data generation clearly outpaces sequence analysis. Therefore, innovative approaches for data handling, visualization and analysis are required.

1.5 Motivation and Research Aims

Sequence data is being generated faster than it can be processed and the divide is getting bigger. This poses new challenges for its handling and interpretation. Tight integration of contextual data, especially environmental parameters, is essential for improving ecological interpretation of genomic data.
This work addressed the current integration needs of marine ecological genomics in a two fold way (Figure 1.3). First, a bioinformatic platform was developed to integrate sequence and contextual data (Section 2.2). Second, this platform was used for three ecogenomic analysis of marine microbes. The adaption of microbes to fluctuations in their environment was tested (Section 2.3). New hypotheses about protein domains of unknown function were proposed2.4. Planktonic communities were compared using molecular methods and genomic data2.5).

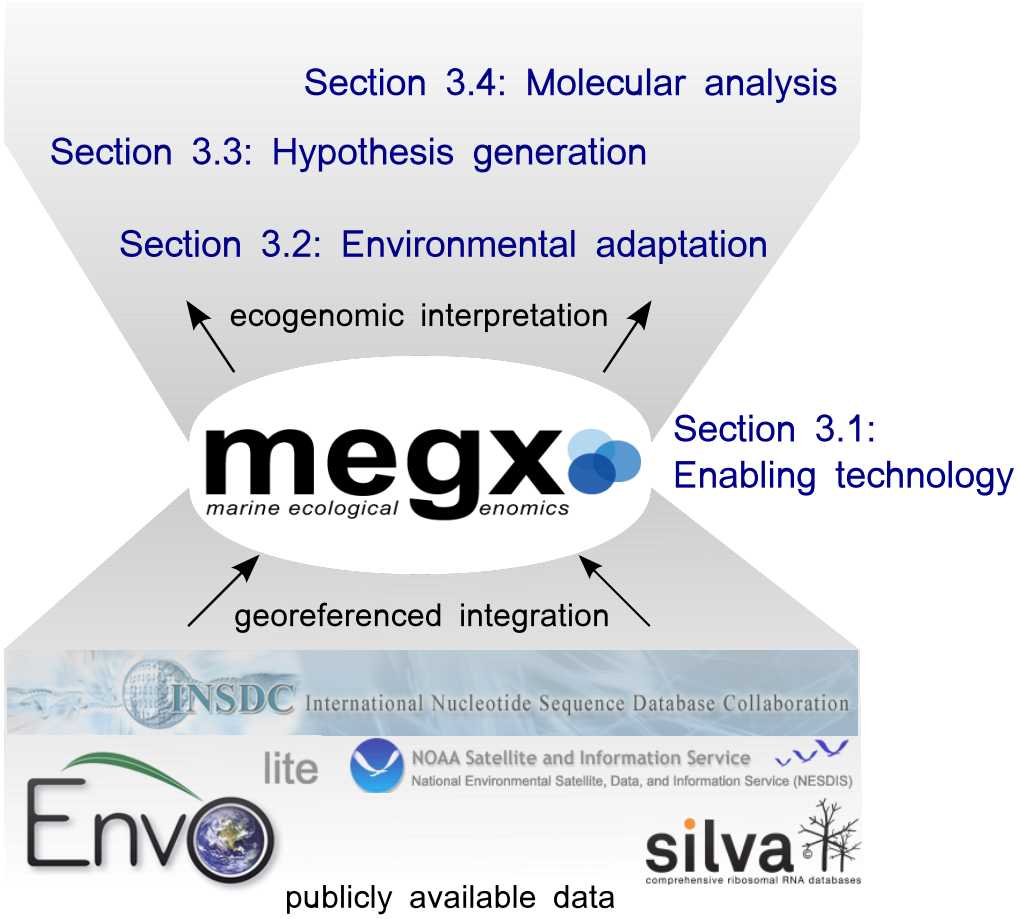


Figure 1.3: An overview of the contribution of this thesis to the advancement of marine ecological genomics.

CHAPTER 2

RESULTS AND DISCUSSION

2.1 Overview

This chapter presents the four research articles that best illustrate the achievements of this thesis in regard to the Research Aims discussed in section 1.5. A short overview of these publications follows.

1. **Megx.net: integrated database resource for marine ecological genomics**
   Authors: Renzo Kottmann, Ivalyo Kostadinov, Melissa Beth Duhaime, Pier Luigi Buttigieg, Pelin Yilmaz, Wolfgang Hankeln, Jost Waldmann and Frank Oliver Glöckner
   Published in: Nucleic Acids Research, 2010 Database Issue
   Contribution: database, web page, Geographic-BLAST, data integration, the first two authors contributed equally to this work
   Relevance: Describes the improvement of the Megx.net portal, whose main task is to provide a geo-referenced integration of sequence and environmental data. The Megx.net project provides infrastructure and analysis tools for marine ecological genomics.

2. **Quantifying the Effect of Environment Stability on the Transcription Factor Repertoire of Marine Microbes**
   Authors: Ivaylo Kostadinov, Renzo Kottmann, Alban Ramette, Jost Waldmann, Pier Luigi Buttigieg, Frank Oliver Glöckner
   Submitted to: Microbial Informatics and Experimentation
   Contribution: designed the study (with RK), carried out all analysis, wrote the manuscript
   Relevance: An ecological genomics study using interpolated environmental data from megx.net. Exemplifies the use of integrated

metadata for ecological analyses.

3. **Ecological perspectives on domains of unknown function: a marine point of view**
   Authors: Pier Luigi Buttigieg, Wolfgang Hankeln, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Melissa Beth Duhaime, and Frank Oliver Glöckner
   Submitted to: The ISME Journal
   Contribution: computational protein domain frequencies using Hidden Markov Models
   Relevance: Generating hypothesis about the possible functions of protein domains based on their co-occurrence patterns and environmental gradients.

4. **Phylogenetic Characterisation of Picoplanktonic Populations with High and Low Nucleic Acid Content in the North Atlantic Ocean**
   Authors: Martha Schattenhofer, Jörg Wulf, Ivalyo Kostadinov, Frank Oliver Glöckner, Mikhail V. Zubkov, Bernhard M. Fuchs
   Published in: Systematic and Applied Microbiology, in press
   Contribution: genomic data collection and integration
   Relevance: A classic molecular ecology study of bacterial plankton. Integrated metadata from megx.net was used.

2.2 Megx.net: integrated database resource for marine ecological genomics

Authors: Renzo Kottmann, Ivalyo Kostadinov, Melissa Beth Duhaime, Pier Luigi Buttigieg, Pelin Yilmaz, Wolfgang Hankeln, Jost Waldmann and Frank Oliver Glöckner

Published in: Nucleic Acids Research, 2010 Database Issue

Contribution: database, web page, Geographic-BLAST, data integration, the first two authors contributed equally to this work

Relevance: Describes the improvement of the Megx.net portal, whose main taks is to provide a geo-referenced integration of sequence and environmental data. The Megx.net project provides infrastructure and analysis tools for marine ecological genomics.

Published online 25 October 2009 *Nucleic Acids Research, 2010, Vol. 38, Database issue* ***D391–D395***
doi:10.1093/nar/gkp918

Megx.net: integrated database resource for marine ecological genomics

Renzo Kottmann[1,*], Ivalyo Kostadinov[1,2], Melissa Beth Duhaime[1,2], Pier Luigi Buttigieg[1,2], Pelin Yilmaz[1,2], Wolfgang Hankeln[1,2], Jost Waldmann[1] and Frank Oliver Glöckner[1,2]

[1]Microbial Genomics Group, Max Planck Institute for Marine Microbiology, D-28359 Bremen and [2]Jacobs University Bremen gGmbH, D-28759 Bremen, Germany

Received September 15, 2009; Accepted October 8, 2009

ABSTRACT

Megx.net is a database and portal that provides integrated access to georeferenced marker genes, environment data and marine genome and metagenome projects for microbial ecological genomics. All data are stored in the Microbial Ecological Genomics DataBase (MegDB), which is subdivided to hold both sequence and habitat data and global environmental data layers. The extended system provides access to several hundreds of genomes and metagenomes from prokaryotes and phages, as well as over a million small and large subunit ribosomal RNA sequences. With the refined Genes Mapserver, all data can be interactively visualized on a world map and statistics describing environmental parameters can be calculated. Sequence entries have been curated to comply with the proposed minimal standards for genomes and metagenomes (MIGS/MIMS) of the Genomic Standards Consortium. Access to data is facilitated by Web Services. The updated megx.net portal offers microbial ecologists greatly enhanced database content, and new features and tools for data analysis, all of which are freely accessible from our webpage http://www.megx.net.

INTRODUCTION

Over the last years, molecular biology has undergone a paradigm shift, moving from a single experiment science to a high-throughput endeavour. Although the genomic revolution is rooted in medicine and biotechnology, it is currently the environmental sector, specifically the marine, which delivers the greatest quantity of data. Marine ecosystems, covering >70% of the Earth's surface, host the majority of biomass and significantly contribute to global organic matter and energy cycling. Microorganisms are known to be the 'gatekeepers' of these processes and insights into their lifestyle and fitness will enhance our ability to monitor, model and predict future changes.

Recent developments in sequencing technology have made routine sequencing of whole microbial communities from natural environments possible. Prominent examples in the marine field are the ongoing Global Ocean Sampling (GOS) campaign (1,2) and Gordon and Betty Moore Foundation Marine Microbial Genome Sequencing Project (http://www.moore.org/microgenome/). Notably, the GOS resulted in a major input of new sequence data with unprecedented functional diversity (3). The resulting flood of sequence data available in public databases is an extraordinary resource with which to explore microbial diversity and metabolic functions at the molecular level.

These large-scale sequencing projects bring new challenges to data management and software tools for assembly, gene prediction and annotation—fundamental steps in genomic analysis. Several new dedicated database resources have recently emerged to tackle the current need for large-scale metagenomic data management, namely CAMERA (4), IMG/M (5) and MG-RAST (6).

Nevertheless, it is increasingly apparent that the full potential of comparative genome and metagenome analysis can be achieved only if the geographic and environmental context of the sequence data is considered (7,8). The metadata describing a sample's geographic location and habitat, the details of its processing, from the time of sampling to sequencing and subsequent analyses are important, e.g. modelling species' responses to environmental change or the spread and niche adaptation of bacteria and viruses. This suite of metadata is collectively referred as contextual data (9).

Megx.net is the first database to integrate curated contextual data with their respective genes, genomes and metagenomes in the marine environment (10). Now, the

*To whom correspondence should be addressed. Tel: +49 421 2028974; Fax: +49 421 2028580; Email: rkottman@mpi-bremen.de

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

D392 Nucleic Acids Research, 2010, Vol. 38, Database issue

extended megx.net database resource allows post factum retrieval of interpolated environmental parameters, such as temperature, nitrate, phosphate, etc. for any location in the ocean waters based on profile and remote sensing data. Furthermore, the content has been significantly updated to include prokaryote and marine phage genomes, metagenomes from the GOS project (2) and all georeferenced small and large subunit ribosomal RNA (rRNA) sequences from the SILVA database project (11).

The extended megx.net portal is the first resource of its kind to offer access to this unique combination of data, including manually curated habitat descriptors for genomes, metagenomes and marker genes, their respective contextual data and additionally integrated environmental data. See the megx.net online video tutorial for a guided introduction and overview at http://www.megx.net/portal/tutorial.html (Supplementary Data).

NEW DATABASE STRUCTURE AND CONTENT

The Microbial Ecological Genomics DataBase (MegDB), the backbone of megx.net, is a centralized database based on the PostgreSQL database management system. The georeferenced data concerning geographic coordinates and time are managed with the PostGIS extension to PostgreSQL. PostGIS implements the 'Simple Features Specification for SQL' standard recommended by the Open Geospatial Consortium (OGC; http://www.opengeospatial.org/), and therefore offers hundreds of geospatial manipulation functions.

MegDB is comprised of (i) MetaStorage, which stores georeferenced DNA sequence data from a collection of genomes, metagenomes and genes of molecular environmental surveys, with their contextual data, and (ii) OceaniaDB, which stores georeferenced quantitative environmental data (Figure 1).

Contextual and sequence data content

Sequences in MetaStorage are retrieved from the International Nucleotide Sequence Database Collaboration (INSDC, http://www.insdc.org/). However, as of September 2009, GOLD reported 5776 genome projects, of which, only 1095 were finished and published (http://www.genomesonline.org/gold.cgi). As most of the sequenced functional diversity is contained in these draft and shotgun datasets, megx.net was extended to host draft genomes and whole genome shotgun data. Currently, MegDB contains 1832 prokaryote genomes (940 incomplete or draft) and 80 marine shotgun metagenomes from the GOS microbial dataset. Marine viruses are a missing link in the correlation of microbial sequence data with contextual information to elucidate diversity and function. Consequently, megx.net now incorporates all sequenced marine phage genomes in MegDB, the first step towards a community call for integration of viral genomic and biogeochemical data (12).

In an effort towards integrating microbial diversity with specific sampling sites, megx.net has been extended to include georeferenced small and large subunit rRNA sequences from the SILVA rRNA databases project

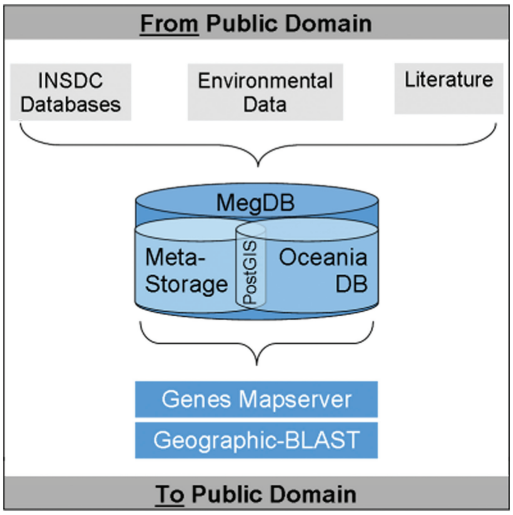


Figure 1. General architecture of megx.net: DNA sequence data (from INSDC) is integrated with contextual data from diverse resources (i.e. manual literature mining and the GOLD database) and interpolated environmental data. MegDB integrates the data conforming to OGC standards and MIGS/MIMS specification. The core megx.net tools, Genes Mapserver and Geographic-BLAST access the MegDB content.

(11). Currently, only 9% (16S/18S) and 2% (23S/28S) of over 1 million sequences in SILVA SSUParc (16S/18S) and LSUParc (23S/28S) databases are georeferenced. With the implementation of the Minimal Information about an Environmental Sequence (MIENS) standard for marker gene sequences (http://gensc.org/gc_wiki/index.php/MIENS), efforts are ongoing to significantly improve this situation.

All genomic sequences in megx.net are supplemented by contextual data from GOLD (13) and NCBI Genome Projects (http://www.ncbi.nlm.nih.gov/genomes/MICROBES/microbial_taxtree.html). The database is designed to store all contextual data recommended by the Genomics Standards Consortium, and is thus compliant with the Minimum Information about a Genome Sequence (MIGS) standard and its extension, Minimum Information about a Metagenome Sequence (MIMS) (7,9).

Furthermore, megx.net is the first resource to provide a manually annotated collection of genomes using terms from EnvO-Lite (Rev. 1.4), a subset of the Environment Ontology (EnvO) (14). An EnvO-Lite term was assigned to each genome project, identifying the environment where its original sample material was obtained. The annotation can be browsed on the megx.net portal using, e.g. tag clouds, and may be used as a categorical variable in comparative analyses.

Environmental data content

OceaniaDB was added to MegDB to supplement the georeferenced molecular data of MetaStorage with interpolated environmental parameters. When sufficient date, depth and location measurements are provided, any 'on site' contextual data taken at a sampling site can

be supplemented by environmental data describing physical, chemical, geological and biological parameters, such as ocean water temperature and salinity, nutrient concentrations, organic matter and chlorophyll.

The environmental data is retrieved from three sources:

(1) World Ocean Atlas: a set of objectively analysed (one decimal degree spatial resolution) climatological fields of *in situ* measurements (http://www.nodc.noaa.gov/OC5/WOA05/pr_woa05.html);
(2) World Ocean Database: a collection of scientific, quality-controlled ocean profiles (http://www.nodc.noaa.gov/OC5/WOD05/pr_wod05.html); and
(3) SeaWIFS chlorophyll *a* data (http://seawifs.gsfc.nasa.gov).

These data are described at 33 standard depths for annual, seasonal and monthly intervals. Together, the location and time data (x, y, z and t) serve as a universal anchor, and link environmental data to the sequence and contextual data in MetaStorage (Figure 1). As such, megx.net integrates biologist-supplied sequence and contextual data (measured at the time of sampling) with oceanographic data provided by third-party databases. All environmental data are compatible with OGC standards (http://www.opengeospatial.org/standards) and are described with exhaustive meta-information consistent with the ISO 19115 standard.

Moreover, based on the integrated environmental data, megx.net provides information to aid biologists in grasping the ocean stability, on both global and local scales. For all environmental parameters, the yearly standard deviations of the monthly values can be viewed on a world map, for easy visualization of high and low variation sample sites. Furthermore, for each sample site, users can view trends in numerous parameters.

USER ACCESS

Genes Mapserver

The Genes Mapserver (formerly Metagenomes Mapserver) offers a sample-centric view of the georeferenced MetaStorage content. Substantial improvements to the underlying Geographic Information System (GIS) and web view have been made. The website is now interactive, offering user-friendly navigation and an overlay of the OceaniaDB environmental data layers to display sampling sites on a world map in their environmental context. Sample site details and interpolated data can be retrieved by clicking the sampling points on the map (Figure 2).

The GIS Tools of the Genes Mapserver allow extraction of interpolated values for several physicochemical and biological parameters, such as temperature, dissolved oxygen, nitrate and chlorophyll concentrations, over specified monthly, seasonally or annually intervals (Figure 2f).

Geographic-BLAST

The Geographic-BLAST tool queries the MegDB genome, metagenome, marine phages and rRNA sequence data using the BLAST algorithm (15). The results are reported according to the sample locations (when provided) of the database hits. With the updated Geographic-BLAST, results are plotted on the Genes Mapserver world map, where they are labeled by number of hits per site (Figure 2). Standard BLAST results are shown in a table, which also provides direct access to the associated contextual data of the hits.

Software extensions to the portal

In addition to the services directly provided by megx.net, the project serves as a portal to software for general data analysis in microbial genomics.

MetaBar (http://www.megx.net/metabar) is a tool developed with the aim to help investigators efficiently capture, store and submit contextual data gathered in the field. It is designed to support the complete workflow from the sampling event up to the metadata-enriched sequence submission to an INSDC database.

MicHanThi (http://www.megx.net/michanthi) is a software tool designed to facilitate the genome annotation process through rapid, high-quality prediction of gene functions. It clearly out-performs the human annotator in terms of accuracy and reproducibility.

JCoast [http://www.megx.net/jcoast; (16)] is a desktop application primarily designed to analyze and compare (meta)genome sequences of prokaryotes. JCoast offers a flexible graphical user interface, as well as an application programming interface that facilitates back-end data access to GenDB projects (17). JCoast offers individual, cross genome and metagenome analysis, including access to Geographic-BLAST.

User test case

To demonstrate the interpretation of genomic content in environmental context, consider a test case with the marine phages. Marine phage genomes (18) and 'viral' classified GOS scaffolds (19) have revealed host-related metabolic genes involved in, i.e. photosynthesis, phosphate stress, antibiotic resistance, nitrogen fixation and vitamin biosynthesis. Geographic-BLAST can be used to investigate the presence of PhoH (accession YP_214558), a phosphate stress response gene, among the sequenced marine phages. The search results can then be interpreted in their environmental context, either as (i) average annual phosphate measurements, or (ii) stability of phosphate concentrations in terms of monthly SD (Figure 2c and d). A closer look at a single genome sample site reveals that *in situ* temperature was not originally reported (Figure 2e), whereas the interpolated data supplements this parameter, among others (Figure 2f).

Web Services

The newly extended version of megx.net offers programmatic access to MegDB content via Web Services, a powerful feature for experienced users and developers. All geographical maps can be retrieved via simple web requests, as specified by the Web Map Service (WMS) standard. The base URL for WMS requests is http://www.megx.net/wms/gms, where more detailed

D394 Nucleic Acids Research, 2010, Vol. 38, Database issue

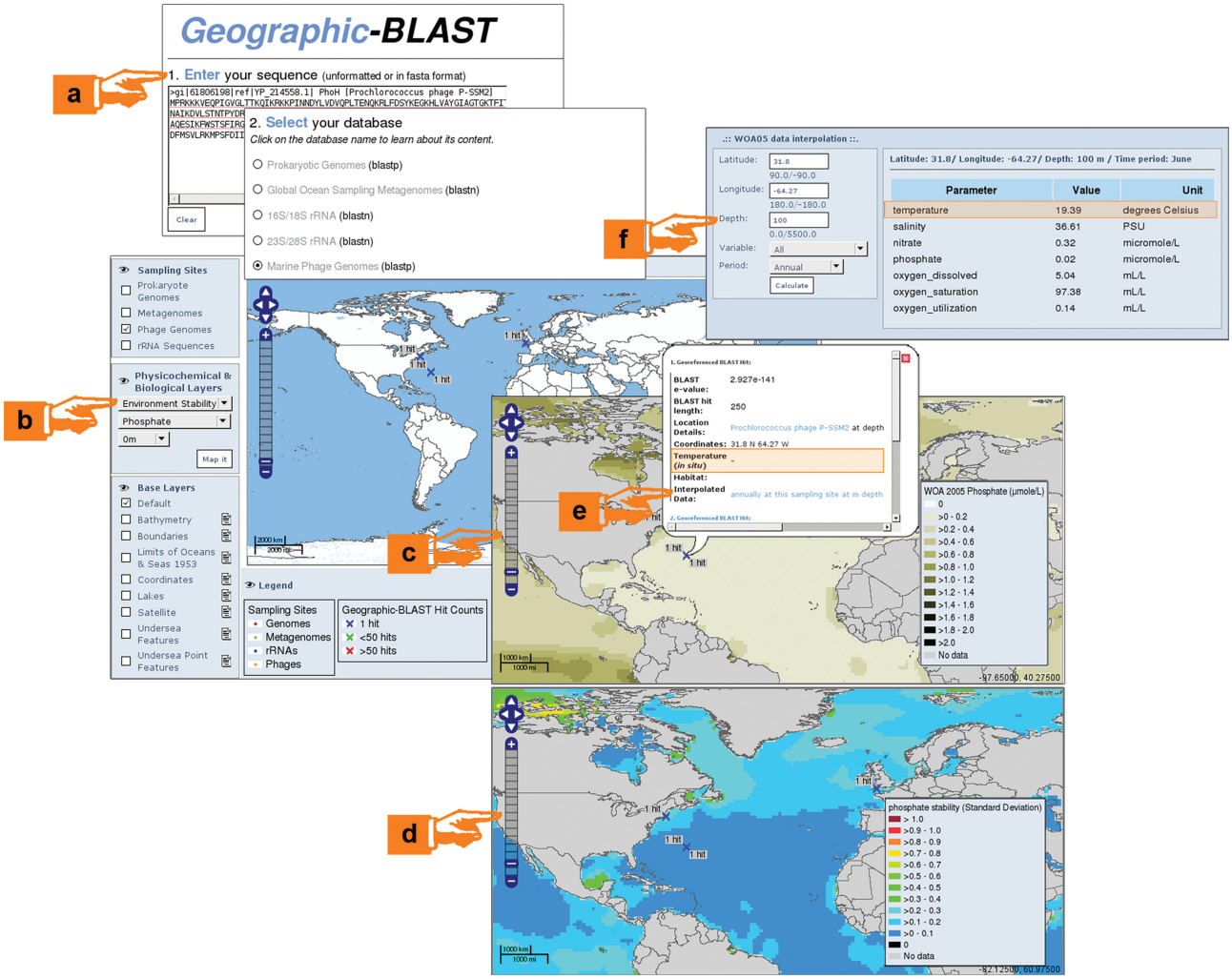


Figure 2. User test case: (**a**) BLAST sequence against the marine phage genomes to see the results on the Genes Mapserver. (**b**) View the BLAST hits with underlying environmental data, such as (**c**) average annual phosphate values, or (**d**) stability of phosphate concentrations in terms of monthly standard deviations. (**e**) BLAST result information can be displayed in a pop-up window, (**f**) where you can link out to megx.net's GIS data interpolator.

information on how to use this service can be found. Megx.net also provides access to MIGS/MIMS reports in Genomic Contextual Data Markup Language (GCDML) XML files for all marine phage genomes through similar HTTP queries, e.g. http://www.megx.net/gcdml/Prochlorococcus_phage_P-SSP7.xml (7,9).

Other changes

The massive influx of sequence data in the last years will out-compete the ability of scientists to analyze it (20). This development already pushes megx.net's capability to provide comprehensive pre-computed data to the limit. To better focus on integration of molecular sequence, contextual and environmental data, megx.net no longer offers pre-computed analyses, especially considering that other facilities, such as MG-RAST and CAMERA have emerged. Furthermore, the 'EasyGenomes Browser' has been replaced with links to the NCBI Genome Projects.

SUMMARY

Since its first publication (10), megx.net has undergone extensive development. The web design has been revamped for better user experience, and the database content greatly enhanced, providing considerably more genomes and metagenomes, marine phages and rRNA sequence data.

Megx.net's unique integration of environmental and sequence data allows microbial ecologists and marine scientists to better contextualize and compare biological data, using, e.g. the Genes Mapserver and GIS Tools. The integrated datasets facilitate a holistic approach to understanding the complex interplay between organisms, genes and their environment. As such, megx.net serves as a fundamental resource in the emerging field of ecosystem biology, and paves the road to a better understanding of the complex responses and adaptations of organisms to environmental change.

Database access

The database and all described resources are freely available at http://www.megx.net/.

Continuously updated statistics of the content are available at http://www.megx.net/content. A web feed for news related to megx.net is available at http://www.megx.net/portal/news/. Feedback and comments, the most effective springboard for further improvements, are welcome at http://www.megx.net/portal/contact.html and via email to megx@mpi-bremen.de.

Overall, it is important to note that the megx.net website does not fully reflect the content and search functionalities of MegDB. For any specialized data request, contact the corresponding author.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online.

ACKNOWLEDGEMENTS

We would like to acknowledge Timmy Schweer, Thierry Lombardot, Magdalena Golden and Laura Sandrine for their valuable input to megx.net, as well as David E. Todd for redesigning the web page.

FUNDING

FP6 EU project MetaFunctions (CT 511784); Network of Excellence 'Marine Genomics Europe'; Max Planck Society. Funding for open access charge: Max Planck Society.

Conflict of interest statement. None declared.

REFERENCES

1. Venter,J.C., Remington,K., Heidelberg,J.F., Halpern,A.L., Rusch,D., Eisen,J.A., Wu,D.Y., Paulsen,I., Nelson,K.E., Nelson,W. *et al.* (2004) Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, **304**, 66–74.
2. Rusch,D.B., Halpern,A.L., Sutton,G., Heidelberg,K.B., Williamson,S., Yooseph,S., Wu,D., Eisen,J.A., Hoffman,J.M., Remington,K. *et al.* (2007) The Sorcerer II Global ocean sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol.*, **5**, e77.
3. Yooseph,S., Sutton,G., Rusch,D.B., Halpern,A.L., Williamson,S.J., Remington,K., Eisen,J.A., Heidelberg,K.B., Manning,G., Li,W. *et al.* (2007) The Sorcerer II Global Ocean Sampling expedition: expanding the universe of protein families. *PLoS Biol.*, **5**, e16.
4. Seshadri,R., Kravitz,S.A., Smarr,L., Gilna,P. and Frazier,M. (2007) CAMERA: a community resource for metagenomics. *PLoS Biol.*, **5**, e75.
5. Markowitz,V.M., Ivanova,N.N., Szeto,E., Palaniappan,K., Chu,K., Dalevi,D., Chen,I.M.A., Grechkin,Y., Dubchak,I., Anderson,I. *et al.* (2008) IMG/M: a data management and analysis system for metagenomes. *Nucleic Acid Res.*, **36**, D534–D538.
6. Meyer,F., Paarmann,D., D'Souza,M., Olson,R., Glass,E.M., Kubal,M., Paczian,T., Rodriguez,A., Stevens,R., Wilke,A. *et al.* (2008) The Metagenomics RAST server–a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, **9**, 386.
7. Field,D., Garrity,G., Gray,T., Morrison,N., Selengut,J., Sterk,P., Tatusova,T., Thomson,N., Allen,M.J., Angiuoli,S.V. *et al.* (2008) The minimum information about a genome sequence (MIGS) specification. *Nat. Biotechnol.*, **26**, 541–547.
8. Field,D., Morrison,N., Glöckner,F.O., Kottmann,R., Cochrane,G., Vaughan,R., Garrity,G., Cole,J., Hirschman,L., Schriml,L. *et al.* (2008) Working together to put molecules on the map. *Nature*, **453**, 978.
9. Kottmann,R., Gray,T., Murphy,S., Kagan,L., Kravitz,S., Lombardot,T., Field,D., Glöckner,F.O. and Genomic Standards Consortium. (2008) A standard MIGS/MIMS compliant XML schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *OMICS*, **12**, 115–121.
10. Lombardot,T., Kottmann,R., Pfeffer,H., Richter,M., Teeling,H., Quast,C. and Glöckner,F.O. (2006) Megx.net–database resource for marine ecological genomics. *Nucleic Acid Res.*, **34**, D390–D393.
11. Pruesse,E., Quast,C., Knittel,K., Fuchs,B.M., Ludwig,W.G., Peplies,J. and Glöckner,F.O. (2007) SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *Nucleic Acid Res.*, **35**, 7188–7196.
12. Brussaard,C.P.D., Wilhelm,S.W., Thingstad,F., Weinbauer,M.G., Bratbak,G., Heldal,M., Kimmance,S.A., Middelboe,M., Nagasaki,K., Paul,J.H. *et al.* (2008) Global-scale processes with a nanoscale drive: the role of marine viruses. *ISME J.*, **2**, 575–578.
13. Liolios,K., Mavromatis,K., Tavernarakis,N. and Kyrpides,N.C. (2008) The Genomes On Line Database (GOLD) in 2007: status of genomic and metagenomic projects and their associated metadata. *Nucleic Acid Res.*, **36**, D475–D479.
14. Hirschman,L., Clark,C., Cohen,K.B., Mardis,S., Luciano,J., Kottmann,R., Cole,J., Markowitz,V., Kyrpides,N., Morrison,N. *et al.* (2008) Habitat-Lite: a GSC case study based on free text terms for environmental metadata. *OMICS*, **12**, 129–136.
15. Altschul,S.F., Gish,W., Miller,W., Myers,E.W. and Lipman,D.J. (1990) Basic local alignment search tool. *J. Mol. Biol.*, **215**, 403–410.
16. Richter,M., Lombardot,T., Kostadinov,I., Kottmann,R., Duhaime,M.B., Peplies,J. and Glöckner,F.O. (2008) JCoast - a biologist-centric software tool for data mining and comparison of prokaryotic (meta) genomes. *BMC Bioinformatics*, **9**, 177.
17. Meyer,F., Goesmann,A., McHardy,A.C., Bartels,D., Bekel,T., Clausen,J., Kalinowski,J., Linke,B., Rupp,O., Giegerich,R. *et al.* (2003) GenDB–an open source genome annotation system for prokaryote genomes. *Nucleic Acid Res.*, **31**, 2187–2195.
18. Sullivan,M.B., Coleman,M.L., Weigele,P., Rohwer,F. and Chisholm,S.W. (2005) Three *Prochlorococcus* cyanophage genomes: signature features and ecological interpretations. *PLoS Biol.*, **3**, 790–806.
19. Williamson,S.J., Rusch,D.B., Yooseph,S., Halpern,A.L., Heidelberg,K.B., Glass,J.I., Andrews-Pfannkoch,C., Fadrosh,D., Miller,C.S., Sutton,G. *et al.* (2008) The Sorcerer II Global Ocean Sampling Expedition: metagenomic characterization of viruses within aquatic microbial samples. *PLoS ONE*, **3**, e1456.
20. (2009) Metagenomics versus Moore's law. *Nat. Methods*, **6**, 623.

2.3 Quantifying the Effect of Environment Stability on the Transcription Factor Repertoire of Marine Microbes

Authors: Ivaylo Kostadinov, Renzo Kottmann, Alban Ramette, Jost Waldmann, Pier Luigi Buttigieg, Frank Oliver Glöckner
Submitted to: Microbial Informatics and Experimentation
Contribution: designed the study (with RK), carried out all analysis, wrote the manuscript
Relevance: An ecological genomics study using interpolated environmental data from megx.net. Exemplifies the use of integrated metadata for ecological analyses.

Quantifying the Effect of Environment Stability on the Transcription Factor Repertoire of Marine Microbes

Ivaylo Kostadinov[1,2], Renzo Kottmann[1], Alban Ramette[1], Jost Waldmann[1], Pier Luigi Buttigieg[1,2], Frank Oliver Glöckner[1,2,§]

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

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

[§] Corresponding author

Email addresses:

IK: ikostadi@mpi-bremen.de

RK: rkottman@mpi-bremen.de

AR: aramette@mpi-bremen.de

JW: jwaldman@mpi-bremen.de

PLB: pbuttigi@mpi-bremen.de

FOG: fog@mpi-bremen.de

Abstract

Background

DNA-binding transcription factors (TFs) regulate cellular functions in prokaryotes, often in response to environmental stimuli. Thus, the environment exerts constant selective pressure on the TF gene content of microbial communities. Recently a study on marine *Synechococcus* strains detected differences in their genomic TF content related to environmental adaptation, but so far the effect of environmental parameters on the content of TFs in bacterial communities has not been systematically investigated.

Results

We quantified the effect of environment stability on the transcription factor repertoire of marine pelagic microbes from the Global Ocean Sampling (GOS) metagenome using interpolated physico-chemical parameters and multivariate statistics. Thirty-five percent of the variation in total TF content could be explained by environment stability. Six percent was attributable to space but none to a combination of both space and stability. Some individual TFs showed a stronger relationship to environment stability and space than the total TF pool.

Conclusions

Environmental stability appears to have a clearly detectable effect on TF gene content in bacterioplanktonic communities described by the GOS metagenome. Interpolated environmental parameters were shown to compare well to *in situ* measurements and were essential for quantifying the effect of the environment on the TF content. It is demonstrated that comprehensive and well-structured contextual data will strongly enhance our ability to interpret the functional potential of microbes from metagenomic data.

Keywords

transcription factors, ecological metagenomics, interpolated environmental data, multivariate statistics

Background

Microorganisms constantly adapt to their environment to survive. An efficient response mechanism is the regulation of transcription, the first step in gene expression, according to environmental demands. Transcription factors (TFs) are the primary agents that perform transcriptional regulation [1]. They consist of a DNA-binding domain (DBD) that typically targets regulatory elements upstream of a gene and an effector domain [2]. The majority of TFs operate by influencing the downstream transcription process and can be classified into 10 super-families according to their DNA-binding mechanisms [3]. Based on the number of genes they regulate, TFs can be divided into 'global regulators' and 'fine tuners' [4]. Both types exert targeted control over gene expression. Global regulators affect a larger number of genes from diverse metabolic pathways and respond to a wider set of stimuli [4, 5]. Conversely, fine tuners are triggered by more specific stimuli and control fewer genes. Up to 10% of bacterial gene products may be devoted to gene regulation [6], a proportion supported by *in silico* analysis of TF abundance in 123 bacterial and archaeal genomes [7]. Although the maximum number of TFs in prokaryotic genomes is bound by the degrees of freedom in their binding mechanisms, larger genomes tend to have more TFs [1, 3]. A greater number of TFs may enable more precise control of gene expression which is required by a complex lifestyle [6]. In general, free-living *Bacteria* and *Archaea* from dynamic environments possess more TFs than those from stable environments [8]. Recently, the effect of environmental factors on gene expression has been studied in the marine model organism *Rhodopirellula baltica* $SH1^T$ [9]. Although only 2% of its gene content is dedicated to transcriptional control [10], it showed an apt regulation response to environmental stress.

Palenik and co-workers (2006) reported that the gene content of two marine *Synechococcus* strains, one isolated from coastal waters and the other from the open ocean, reflect the variability of their respective environments [11]. The coastal strain possessed a higher number of sensors and response regulators when compared to the open ocean strain, allowing it to respond to its dynamic environment. Gianoulis and coworkers (2009) investigated the environmental adaptation of

metabolic pathways in the Global Ocean Sampling (GOS) metagenomes [12]. They observed no significant differences in the abundance of transcriptional/translational pathways between these two groups of samples, loosely described as open ocean and coastal. A more recent study described environmental adaptation in 197 marine microbial genomes and related the findings to the GOS metagenome [13]. The abundant cosmopolitan species which are adapted to slow growth in nutrient-poor conditions have a smaller genome size, lower metabolic plasticity, and fewer transcriptional regulators than their counterparts which are adapted to alternating periods of 'feast and famine'. However, quantifying the effect of the environment on the transcription factor repertoire of marine microbes remains a challenge. A comprehensive set of environmental parameters, describing the samples at the time they were taken and the sampling location over monthly to yearly time scales, is a prerequisite for addressing this question. Unfortunately, environmental *in situ* measurements taken during sampling are often missing or incomplete. Even when they are at hand, they give only a static 'snapshot' of the environmental conditions. The use of interpolated parameters can help to overcome these shortcomings: they can replace missing values, describe sampling sites in different temporal scales and give indications of the stability of the environment. A few metagenomic studies have taken advantage of these features of interpolated environmental parameters. Gianoulis and coworkers (2009) validated imputed salinity values against extrapolations from the World Ocean Database [14]. Rusch and coworkers (2010) used monthly averages for nitrate and phosphate from the World Ocean Atlas (WOA) to study the *Prochlorococcus* clades detected in the GOS metagenome with respect to nutrient availability [15].

Here we investigated the influence of environment stability on TF gene content in the GOS metagenome [16, 17]. To this end, we (1) compared interpolated environmental parameters against on-site measurements to verify the predictive power of the interpolations used; (2) calculated a yearly stability measure for each environmental parameter based on 12 monthly averages; (3) applied an alternative method to standardize the metagenomic samples for size in order to make

protein domain counts comparable; (4) applied redundancy analysis (RDA) to assess the effect of environmental stability and space on the TF content; (5) used multiple linear regression (MLR) to identify possible dependencies between single TFs, combinations of stability parameters, and space.

Results and Discussion

Interpolated environment parameters compare well to *in situ* measurements

We selected GOS samples where on-site measurements and monthly interpolated values for temperature (55 samples) and salinity (44 samples) were available. We calculated a linear regression model using interpolated monthly parameter values to predict values measured on board the Sorcerer II during sampling. Both interpolated temperature and salinity values proved to be good estimators of the measured values, with a Pearson correlation coefficient (R^2) of 0.76 (p-value < 2.2e-16) and 0.6 respectively (p-value = 2.459e-10) (Figure 1). Coastal areas, however, pose a significant problem for interpolation due to lack of reliable data or major terrestrial influences on the water bodies that are hard to quantify (e.g. riverine input, anthropogenic activity). Sample GS033 came from a hypersaline mangrove forest, an environment that differs markedly from the surrounding water masses. The interpolated monthly average for this sample was 29 Practical Salinity Units (PSU) lower than the measured one. Considering that the area is known to be hypersaline, this large difference is more likely due to an insufficient number of data points available for interpolation rather than by a temporary event taking place at the time of sampling. Supporting this assumption, the interpolated monthly temperature was 12°C lower than the *in situ* measurement. Because no reliable interpolations were possible for GS033, it was excluded from the regression analysis of salinity and from the environment stability analysis. The combination of numerical data with categorical description (hypersaline) of the habitat helped to detect and explain differences between interpolated and *in situ* values. The interpolations for the remaining locations are based on a number of previous *in situ* measurements [18] and easily accessible surface waters, i.e. the first 30 m of the marine epipelagic zone, are well sampled in this regard. This is the probable

reason for the good fit between measured and interpolated monthly values. Our results suggest that numeric interpolation of environmental parameters can complement or, when necessary, even substitute parameters measured *in situ*. These comprehensive datasets can then be used, with a fair degree of confidence, in deriving more complex descriptors of the environment such as its stability.

Variation in single-copy gene numbers

Single-copy genes (SCGs) are genes which are assumed to appear only once per genome. Their total number is suggested to reflect the genome equivalents in metagenomic samples [19]. Therefore, they are good candidates to standardize results of sequence-based searches in samples of different sizes. A very basic approach would be to divide the absolute counts of a TF by the absolute count of an SCG (Formula 1). However, we expected significant differences in the occurrences of different SCGs. To test this assumption, we compared the abundance of 53 prokaryotic SCGs in 58 GOS samples. Four overrepresented and 12 underrepresented SCGs were found (Figure S1.2, Table S1.1). Some of those were outliers in up to 98% of the samples. Over- and under-representation of SCGs was observed in all samples, although the variation dropped with increasing number of sequences per sample (Figure S1.1 and Figure S1.2).

We compared the behavior of basic statistical descriptors like the mean and the median for producing a suitable standardization parameter (Figure S1.3). All descriptors behaved in a similar way, showing an increasing number of SCGs with increasing number of sequences. The interquartile range remained stable regardless of the sample size, showing an almost equal spread of the SCG counts per sample. We performed the analysis of the total TF content using two standardization parameters corresponding to two standard deviations above and below the mean and compared the results. No significant difference was detected, and even if such a difference was observed, using both parameters for calculations would translate into reporting results as a range rather than as a single value.

It is possible that cloning and sequencing biases in the GOS metagenome may explain over- and underrepresentation of certain SCGs. It is also possible that some of the SCGs appear in more than one copy in some genomes. The original work of [20] that identified SCGs was based on 191 completely annotated genomes across the tree of life. At the time of our study, the ENTREZ Genome Project collection (http://www.ncbi.nlm.nih.gov/genomes/lproks.cgi) listed 1446 complete microbial genomes and another 3888 in progress. Furthermore, an EnvO-Lite [21] classification of complete microbial genomes available at the megx.net portal (http://www.megx.net) features 227 marine water column isolates. Given the many-fold increase in microbial genomes, it would be beneficial to re-evaluate the list of SCGs, focusing on marine prokaryotes, but such analysis was beyond the scope of this study. According to [22], the average genome size of a sample and the length of an SCG influence relative counts. The SCGs used here are universally distributed, most of them being related to the translation machinery [20]. Therefore, their presence should be genome-size independent. The effect of gene length on the sampling probability is neutralized by combining the observations from several SCGs with different lengths. Ultimately, we used the mean SCG count per sample as a standardization measure (Formula 2).

The TF content significantly responds to environment stability

We derived eight environment stability measures based on the standard deviation of interpolated monthly temperature, salinity, dissolved oxygen, apparent oxygen utilization (AOU), oxygen saturation, phosphate, nitrate, and silicate measurements over a 12-month period. This was done for 44 of the samples used for the determination of the SCG variation. Because co-varying stability measures may confound statistical analyses, we only retained variables with a correlation coefficient below 0.6 to any other variables (Table S1.2). As expected, nitrate stability correlated strongly with phosphate stability. The tight connection between these two nutrients is well known as the Redfield Ratio [23]. Tyrell (1999) showed the strong correlation between phosphate and nitrogen in the WOA data [24]. Similarly, the amount of dissolved oxygen is known to depend

strongly on water temperature [25]. This relationship showed as a strong correlation ($\rho = 0.75$) between the two stability measures. Oxygen saturation and AOU are both derived from the dissolved oxygen [26] but they showed exceptionally high correlation ($\rho = 0.99$) to each other and moderate correlations to either phosphate ($\rho = 0.63$) or silicate ($\rho = 0.61$). Thus, the stability measures for temperature, salinity, phosphate, and silicate were used for further analysis.

In order to evaluate the effect of the environment stability on the total TF content in 44 GOS samples we used RDA. Combining automatic and manual parameter selection, we found a statistical model in which environment stability and space best described the variation in TF content. The environment stability was represented by temperature stability (p-value < 0.001) and phosphate stability (p-value < 0.1). The environment stability alone accounted for 35% of the variation in TFs. As described above, for pairs of strongly correlating stability measures only one measure was taken; therefore, the effects of two strongly correlating parameters could not be differentiated. Temperature stability could either influence TF variation directly or could indicate the effect of dissolved oxygen stability. The same is true for phosphate stability and nitrate stability. Tyrrell (1999) argues that phosphate limits oceanic primary production on a short time scale, while nitrate limits it on a global time scale [24]. In this study, we cannot speculate on what spatial scale environmental changes cause genomic TF variation in prokaryotes. The space component was represented by one of the two axes (X2), produced by principal coordinate analysis of the Cartesian distances between samples and accounted for 6% of the TF variation (p-value < 0.01). Because many TFs perform universal house-keeping functions, spatial distance alone was expected to explain only a minor proportion of the TF variation. In this case, space could be considered an abstract proxy for the different conditions between spatially separated environments. Contrary to our expectations, no variation could be explained by the combined effect of environment and space in our model. A biplot of the RDA results reveals that the majority of TFs cluster together and the explanatory variables do not have enough discriminatory power (Figure 2). However, several TFs

like aldedh were more strongly affected by the environment stability and space. Overall, 59% of the variation in the TF content remained unexplained and it is clear that further factors are required to explain patterns of TF distribution more completely. Additional environmental parameters and interactions with viruses and eukaryotes are likely to feature among these.

Gianoulis and coworkers (2009) explored the adaptation of metabolic pathways in the GOS metagenome to the environment [12]. They divided the samples in two groups, loosely described as coastal and open ocean. No significant difference in the transcription machinery between two sets was detected. In their estimation, fine-grained relationships between the samples and their environment might have been undetectable by the method used to partition the samples. Although generally similar, our study differs from that of Gianoulis et al. (2009) in several aspects. Their explorative approach was well suited for a broad range of pathways. However, more subtle patterns in specific pathways might remain undetected. Here we focused on one functional group (TFs) and adapted our methods accordingly. We performed the analysis on a six-frame translation of the raw GOS reads to avoid artifacts from assembly and ORF prediction. Further, we used a curated list of Hidden Markov Models (HMM) to detect genes of interest and used an extended set of environmental parameters, including nutrients. Small-scale differences along nutrient gradients are of importance when describing the ecology of microorganisms [27], so we kept the scale as fine-grained as possible. Lastly, we investigated the adaptation of microbial TF repertoire in response to environment stability rather than temporary environmental conditions. We were able to complement the findings of Gianoulis et al. (2009) with a detailed quantification of the TF content adaptation to environmental stability.

A more recent study detected the trend in the TF repertoire of marine microbes we quantified here [13]. The genomes of 197 marine isolates were compared with respect to their coverage in the GOS dataset resulting that only 34 marine genomes are well covered in the GOS dataset. These are very streamlined, having heavily reduced capacities for transcriptional regulation, environment sensing

and amino-acid uptake. The remaining 163 genomes were sparsely covered by the GOS dataset and were more adapted to changing environmental conditions. Yooseph and coworkers concluded that the prevailing picoplankton has a low 'bacterial IQ' [28] and uses alternatives to transcriptional control for metabolic regulation. Our findings from directly querying the metagenome concur with the differences based on trophic strategies observed by Yooseph et al. (2010). With 35% effect of environmental stability on the TF content we have shown that more dynamic environments require different TF repertoires than stable environments.

Single TFs are more tightly connected to environment stability and space

The RDA of total TF content suggested that individual TFs show stronger relationships to environment stability than the total TF content. Using the 44 samples we applied MLR to test the effect of environment stability and space on single TFs. For 19 TFs more than 30% of the variation could be explained by a combination of environmental stability parameters and spatial components (Table 1). Temperature stability was present in all MLR models. Temperature is known to be an important factor in determining bacterial populations and their functions in the oceans [29]. However, temperature might also be a proxy for other parameters. Several TFs were best explained by different combinations of temperature stability, salinity stability and the second spatial axis (X2). Since these factors are rather broad, we inspected more closely the TFs which were co-explained by phosphate (i.e. nutrients) stability and silicate stability.

Nutrient stability co-explained the variability of both broad and specific TFs. Response_reg (PF00072) is a general receptor domain which interacts with a DNA-binding effector domain (often LytTR). The model representing LacI (PF00356) family of regulators is a broad-spectrum DBD. This particular TF was equally well explained by temperature stability and either phosphate or silicate stability. We speculate that this is due to the wide range of regulators belonging to this family. Penicillinase_R (PF03965) is responsible for the repression of the penicillinase gene. Availability of nutrients generally causes increased prokaryotic and eukaryotic cell density in the

water column. The release of beta-lactam antibiotics is a competitive measure in such a scenario which must be met with a well-regulated resistance. In coastal areas, terrestrial input of such antibiotic substances can also be expected. The HTH_6 domain (PF01418) is involved in the regulation of phospho-sugar metabolism, we speculate that we observed a direct link between the function regulated by the TF and the stability of the substrate for this function. Another TF, Trp_repressor (PF01371), regulates the Tryptophan operon and is a classic example for transcription control by attenuation. Tryptophan biosynthesis includes phosphorylated intermediates, so an indirect link between phosphate stability and the distribution of Tryptophan repressors is likely. Additionally, phosphate stability could also be a proxy for the overall nutrient stability, which would influence uptake of amino acids.

Silicate stability co-explained the variation in TFs which describe a scenario where bacterial populations interact with eukaryotes in a dynamic environment. TFs from the HTH_3 family (PF01381) are involved in plasmid copy control and methylation, the latter a means to prevent the digestion of DNA by restriction endonucleases mechanism. TOBE (PF03459) is part of ABC transporters and detection of small ligands like sulphate. LytTR (PF04397) is involved in the control of cell autolysis. Bacterial adaptation includes complex interactions with phytoplankton. Bacterial assemblages mediate silicon regeneration from lysed diatoms, detritus and marine snow [30, 27]. Algal blooms, for example, strongly affect microbial communities [31, 32]. In a bloom situation, precise control over substance detection and transport, defense mechanisms and cell death would provide a selective advantage. Based on the TFs whose variation was co-explained by silicate, we speculate that we have detected a response of bacterial regulatory potential to oscillations in diatom communities, for example during and after an algal bloom.

Our findings on single TFs support the trophic description of the GOS dataset [13]. Typically, copiotrophs are adapted to capitalize on transient nutrient availability on which the survival of their populations strongly depends. They are more influenced by marine eukaryotes (e.g. algal blooms)

and dominate the water column only sporadically [13, 33]. In contrast to microbes with oligotrophic adaptations, copiotrophs still possess the majority of energy uptake systems (e.g. amino acids). The environmental stability effect on the three TFs above supports the general idea of the distinguishing adaptations of copiotrophs and oligotrophs. Additionally, the functions controlled by these TFs might be the key to describing the relationship of copiotrophic communities to their surroundings in greater detail.

Detection limits and interpretation considerations with our approach

The Pfam HMMs [34] used in this study model only key protein domains of the TFs and sometimes represent whole TF families. Therefore, an absolute, one-to-one relationship between a single TF and a particular gene or function is sometimes impossible to infer. Although we used a set of eight environmental parameters, other factors (e.g. predator-prey interactions, viral infections, iron concentration) might significantly contribute to the patterns of TF distribution. Moreover, the interpolated environmental data values were monthly averages which might not reflect smaller temporal variations. These constraints form a certain resolution limit on our findings that is hard to quantify. On the other hand, the selective pressure which the environment stability exerts on bacterial transcription control was strong enough to leave a genomic imprint which is detectable despite this resolution limit. Furthermore, metagenomics provides a glimpse into the genomic potential of microbial communities, but not into their gene expression patterns. Therefore, any dependencies between the environment and the genomic repertoire have to be rather stable. In this study, we focused on linear relationships between TF content and the numeric stability of the environment, but non-linear relationships could also be possible.

Conclusion

Using interpolated environmental data, we detected and quantified an ecogenomic trend in the transcription factor repertoire of marine bacterial communities that depended on spatial distance and

environmental stability. Environment stability was responsible for 35% of the variation in total TF content while 6% was attributed to space. Up to 60% of the variation in single TFs could be attributed to combinations of environment stability factors and space. In several cases the function controlled by the TFs was directly related to the environmental stability measures that best explained their variation. Despite resolution limitations of the data, our results strongly suggest that the effect of environment stability on the genome composition of bacterioplankton is strong enough a detectable signal. Improved availability and integration of contextual data, preferably compliant with the checklists of the Genomics Standards Consortium [35], will make it possible to describe ecogenomic trends with higher resolution and better characterize the influence of the environment on prokaryotic metagenomes.

Methods

Sequence and Environmental Data

Sequence reads and metadata for 82 samples of GOS metagenome were obtained from the Community Cyberinfrastructure for Advanced Microbial Ecology Research & Analysis (CAMERA) website [36]. These include samples from the Sargasso Sea [16], the northwest Atlantic, the eastern tropical Pacific [17], and the Indian Ocean transect. The interpolated environmental data for the GOS samples (Supplement S2.1 and S2.2) was extracted from the portal for Marine Ecological Genomics [37] using the geographic location (based on GPS coordinates), sampling date and depth. The interpolations were based on data from the World Ocean Atlas 2005 [14]. Eight environmental parameters were available, namely temperature, salinity, dissolved oxygen, apparent oxygen utilization (AOU), oxygen saturation, phosphate, nitrate, and silicate.

Ecological modeling

Statistical analyses and plotting were performed using the free software environment for statistical computing and graphics, R [38] with the *vegan* [39], and *MASS* packages [40]. The R code for this study is available in Supplement S2 (Rcode.txt).

For linear regressions of environmental data, all GOS samples where interpolation for temperature and salinity was possible were considered (Supplement S2.1). Only one *in situ* measurement and one interpolated value per sampling site, defined by unique GPS coordinates, time and depth of sampling, are possible. Therefore, only one sample per sampling site was kept. Two samples GS000a and GS000b have the combined sequence content from two different locations (Sargasso Stations 11 and 13) [16]. In this comparison only, GS000a represents the environmental data from Sargasso Station 11 and GS000b that from Sargasso Station 13. Samples where the *in situ* measurement was missing were excluded. This left 55 samples to be compared for temperature and 44 for salinity. The choice of samples for this experiment included no further requirements, because the aim was to demonstrate the accuracy of interpolated data. The interpolations were used as response variables and the *in situ* measurements as explanatory variables. The compared values were expressed in the same units: degrees Celsius for temperature and PSU for salinity. Hence, no further transformation was necessary.

Protein Domain Searches with Hidden Markov Models

The sequence reads of the GOS metagenome were translated in all six reading frames using the transeq tool from the EMBOSS package [41] with default parameters (version 6.1.0). Hidden Markov Models were selected from the Pfam database (release 24) [34]. Unless stated otherwise, descriptions of HMM models and corresponding TF functions were taken from the Pfam website [42]. Protein domain searches were done with HMMER3 in version 3.0b3 using the default parameters [43]. The results were imported into a relational database. Following the "HMMER3 beta test: User's guide" (Version 3.0b3) [44], ***significant*** results were defined by the following criteria: 1) domain independent E-value < 0.001, 2) hmm_to-hmm_from >= 20% of model length and 3) the bias should be at least an order of magnitude smaller than the score.

Single Copy Gene distribution

Samples from GOS were selected to ensure: 1) the filter size used targeted prokaryotes (between 0.1 µm and 0.8 µm) and 2) their origin was not a fresh water environment (based on the habitat type reported in the GOS metadata). Finally, the Sargasso Sea sample GS000a, which is suspected to be contaminated with non-marine *Shewanella* and *Burkholderia* species [45], was removed.

The following samples were excluded from further analysis: GS0 38, 39, 40, 41, 42, 43, 44, 45, 46 and 50. They had extremely low SCG counts, with a maximum per sample average of 1. This was in line with the extremely low number of total sequences in these samples (between 626 and 759 sequences per sample) compared to the rest of the samples (between 11,496 and 692,255 sequences per sample) (Supplement S2.4). A total of 58 samples remained for further analysis (Supplement S2.3). The list of 53 HMMs was based on Ciccarelli et al. 2006 (Table S1.3).

Effect of environment stability on TF content

WOA interpolations were possible for 44 of the 58 GOS samples from the SCG analysis. Additionally, the Mangrove Forest sample GS033 was removed. Environment stability measures is described by the standard deviation of the twelve monthly averages for each interpolated variable at each sampling site (Formula 3, Supplement S2.5). For GS000b, the average from Sargasso Station 11 and 13 was taken. Stability measures were z-scored (Formula 4) to neutralize the effects of different scales and units [46]. Co-varying stability measures were excluded when their Spearman's rank correlation coefficient (ρ) exceeded 0.6 and the test was statistically significant (p-value << 0). The list of TF models was compiled according to Minezaki et al. 2005 [47] (Table S1.4). The list contained 40 DNA-Binding Domains (DBDs) and 26 non-DBDs (Supplement S2.6). The models seemed to be rather stable as only one Pfam HMM model had changed since the time of publication in 2005 (PF02573 was merged into or replaced by PF00126). One of the TF HMMs had no significant hits (CtsR, PF05848) and could not be used for the analysis. The raw counts for each TF HMM in each sample (Supplement S2.7) was standardized using Formula 2 and the mean of the SCG counts for the respective sample.

$$sTF\ count = \frac{raw\ TF\ count}{raw\ SCG\ count}$$

Formula 1: The raw count of each individual TF in a sample is divided by the

$$sTF = \frac{raw\ TF\ count}{f\left(raw\ SCG\ count\right)}$$

Formula 2: The raw count of each individual TF in a sample is divided by a

absolute count of a single SCG to compute a standardized TF (sTF) count, which is independent of the sample size.

value calculated from the absolute counts of several SCGs to compute a standardized TF (sTF) count, which is independent of the sample size.

$$\sigma = \sqrt{\sum_{i=1}^{N} \frac{(X_i - \mu)^2}{N-1}}$$

Formula 3: Sample standard deviation. The individual values (Xi) are monthly interpolated values for one of the eight environmental parameters. In this study, the standard deviation (σ) was used as a stability measure (the lower the SD, the more stable an environment was considered).

$$z = \frac{x - \mu}{\sigma}$$

Formula 4: Z-score transformation. The raw score (x) is transformed by subtracting the population mean (μ) and dividing by the standard deviation (σ). In this study, each stability measure was treated as a raw score across all samples (the population).

Principal coordinate analysis (PCoA) was used to map the spatial components from the Cartesian distances between the samples back to a 2D plane (Supplement S2.8). The distances were calculated from their GPS coordinates, using the geographic information system module of the megx.net relational database MegDb [37]. For GS000b, the average of the distance between the two original samples it incorporates and any other sample was taken. PCoA, also known as metric multidimensional scaling, is an ordination method that can map multidimensional data to fewer dimensions to aid interpretation. In this study, the 2D coordinates of each sample (X1, X2) and polynomial terms (up to third-degree terms) thereof represented the spatial components. RDA,

which is a multivariate extension of linear regression, was used to calculate the effect of environment stability and space on the total TF content. The standardized TF counts were used as response variables and the four environment stability measures (temperature, salinity, phosphate, silicate), the two spatial coordinates (X1, X2) and their associated polynomial terms ($X1^2$, $X1^3$, $X2^2$, $X2^3$) were used as explanatory variables. We applied automatic forward and backward model selection to find the combination of explanatory variables that best explained the variation in the response variables. The combined and independent effect of environment stability and space was tested. The combined model and the independent environmental model both identified temperature stability and phosphate stability as significant explanatory variables. The independent space model identified spatial polynomial terms as significant rather than the X2 from the combined model. We tried to replace X2 in the combined model with combinations of the independent space model; however, no improvement in explained variation or significance levels was observed. Consequently, the combined model was used in further analysis. Variation partitioning was used to separate the effect of environment stability and space. All models and partitions were tested for significance using 1000 permutations of the response data. MLR was used to quantify the effect of environment stability and space on individual TFs. The standardized count of each individual TF per sample was used as a response variable. The explanatory variables were the same as for RDA. We compared different model selection methods based on the Akaike information criterion with 1000 steps. Whenever an automatically generated model explained more than 30% of the variation in a TF ($R^2 > 0.3$), we tried to manually improve it by removing explanatory variables with low significance (p-value > 0.1).

List of Abbreviations

DBD: DNA-binding domain

GOS: Global Ocean Sampling

MLR: Multiple linear regression

RDA: Redundancy analysis

SCG: Single-copy gene

TF: Transcription factor

WOA: World Ocean Atlas

Competing Interests

The authors declare that they have no competing interests.

Authors' contributions

IK and RK conceived and planned the study. IK performed the analysis and drafted the manuscript. AR participated in statistical analysis. JW participated in the Hidden Markov Model searches. PLB participated in the formulation of the study and helped draft the manuscript. FOG coordinated the work and finalized the manuscript. All authors read and approved the final manuscript.

Authors' information

IK is a bioinformatician focusing on functional metagenomics and data integration. RK is a bioinformatician focusing on data integration and standardization. AR is an ecologist with strong background in multivariate statistics. JW is an informatician with strong background in bioinformatic pipeline development. PLB applies a background in biochemistry, cell biology and marine microbiology to the study of ecological genomics. FOG is a Professor of Bioinformatics at Jacobs University and group leader of the Microbial Genomics and Bioinformatics Group at the Max Planck Institute for Marine Microbiology. FOG and RK are active members of the Genomics Standards Consortium.

Acknowledgements

This work was funded by the Max Planck Society.

References

1. Charoensawan V, Wilson D, Teichmann SA: **Genomic repertoires of DNA-binding transcription factors across the tree of life**. *Nucleic Acids Research* 2010, **38**:7364 -7377.
2. Nowick K, Stubbs L: **Lineage-specific transcription factors and the evolution of gene regulatory networks**. *Briefings in Functional Genomics* 2010, **9**:65-78.
3. Itzkovitz S, Tlusty T, Alon U: **Coding limits on the number of transcription factors**. *BMC Genomics* 2006, **7**:239.
4. Babu MM, Luscombe NM, Aravind L, Gerstein M, Teichmann SA: **Structure and evolution of transcriptional regulatory networks**. *Curr. Opin. Struct. Biol* 2004, **14**:283-291.
5. Martínez-Antonio A, Collado-Vides J: **Identifying global regulators in transcriptional regulatory networks in bacteria**. *Curr. Opin. Microbiol* 2003, **6**:482-489.
6. Pérez-Rueda E, Collado-Vides J, Segovia L: **Phylogenetic distribution of DNA-binding transcription factors in bacteria and archaea**. *Computational Biology and Chemistry* 2004, **28**:341-350.
7. Lombardot T, Bauer M, Teeling H, Amann R, Glöckner FO: **The transcriptional regulator pool of the marine bacterium Rhodopirellula baltica SH 1T as revealed by whole genome comparisons**. *FEMS Microbiol. Lett* 2005, **242**:137-145.
8. Cases I, de Lorenzo V, Ouzounis CA: **Transcription regulation and environmental adaptation in bacteria**. *Trends Microbiol* 2003, **11**:248-253.
9. Wecker P, Klockow C, Ellrott A, Quast C, Langhammer P, Harder J, Glöckner FO: **Transcriptional response of the model planctomycete Rhodopirellula baltica SH1(T) to changing environmental conditions**. *BMC Genomics* 2009, **10**:410.
10. Glöckner FO, Kube 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: **Complete genome**

sequence of the marine planctomycete Pirellula sp. strain 1. *Proc Natl Acad Sci U S A* 2003, **100**:8298-8303.

11. Palenik B, Ren Q, Dupont CL, Myers GS, Heidelberg JF, Badger JH, Madupu R, Nelson WC, Brinkac LM, Dodson RJ, Durkin AS, Daugherty SC, Sullivan SA, Khouri H, Mohamoud Y, Halpin R, Paulsen IT: **Genome sequence of Synechococcus CC9311: Insights into adaptation to a coastal environment**. *Proc Natl Acad Sci U S A* 2006, **103**:13555-13559.
12. Gianoulis TA, Raes J, Patel PV, Bjornson R, Korbel JO, Letunic I, Yamada T, Paccanaro A, Jensen LJ, Snyder M, Bork P, Gerstein MB: **Quantifying environmental adaptation of metabolic pathways in metagenomics**. *Proc Natl Acad Sci U S A* 2009, **106**:1374–1379.
13. Yooseph S, Nealson KH, 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 Y, Friedman R, Frazier M, Venter JC: **Genomic and functional adaptation in surface ocean planktonic prokaryotes**. *Nature* 2010, **468**:60-66.
14. Boyer, T. P.: *World Ocean Database 2005*. U.S. Government Printing Office, Washington, D.C.; 2006.
15. Rusch DB, Martiny AC, Dupont CL, Halpern AL, Venter JC: **Characterization of Prochlorococcus clades from iron-depleted oceanic regions**. *Proceedings of the National Academy of Sciences* 2010, **107**:16184-16189.
16. Venter JC, Remington K, Heidelberg JF, Halpern AL, Rusch D, Eisen JA, Wu D, Paulsen I, Nelson KE, Nelson W, Fouts DE, Levy S, Knap AH, Lomas MW, Nealson K, White O, Peterson J, Hoffman J, Parsons R, Baden-Tillson H, Pfannkoch C, Rogers Y, Smith HO: **Environmental Genome Shotgun Sequencing of the Sargasso Sea**. *Science* 2004, **304**:66-74.
17. Rusch DB, Halpern AL, 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 Y, 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: **The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific**. *PLoS Biol* 2007, **5**:e77.

18. **World Ocean Atlas 2005 (WOA05) Product Documentation** [ftp://ftp.nodc.noaa.gov/pub/WOA05/DOC/woa05documentation.pdf].

19. Raes J, Korbel JO, Lercher MJ, von Mering C, Bork P: **Prediction of effective genome size in metagenomic samples**. *Genome Biol* 2007, **8**:R10.

20. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, Bork P: **Toward Automatic Reconstruction of a Highly Resolved Tree of Life**. *Science* 2006, **311**:1283-1287.

21. **The Environment Ontology (EnvO)** [http://www.environmentontology.org/].

22. Beszteri B, Temperton B, Frickenhaus S, Giovannoni SJ: **Average genome size: a potential source of bias in comparative metagenomics**. *ISME J* 2010, **4**:1075-1077.

23. Redfield A: **On the proportions of organic derivatives in sea water and their relation to the composition of plankton**. In *James Johnstone Memorial Volume*. Liverpool University Press, Liverpool, UK; 1934:172-196.

24. Tyrrell T: **The relative influences of nitrogen and phosphorus on oceanic primary production**. *Nature* 1999, **400**:525-531.

25. Weiss R: **The solubility of nitrogen, oxygen and argon in water and seawater**. *Deep Sea Research and Oceanographic Abstracts* 1970, **17**:721-735.

26. H. E. Garcia, Locarnini, R. A., T. P. Boyer, J. I. Antonov: *World Ocean Atlas 2005, Volume 3: Dissolved Oxygen, Apparent Oxygen Utilization, and Oxygen Saturation.* S. Levitus, Ed. NOAA Atlas NESDIS 63, U.S. Government Printing Office, Washington, D.C., 342 pp.; 2006.

27. Azam F, Malfatti F: **Microbial structuring of marine ecosystems**. *Nat Rev Micro* 2007, **5**:782-791.

28. Galperin MY: **A census of membrane-bound and intracellular signal transduction proteins in bacteria: Bacterial IQ, extroverts and introverts**. *BMC Microbiol* 2005, **5**:5-35.

29. Fuhrman JA, Steele JA, Hewson I, Schwalbach MS, Brown MV, Green JL, Brown JH: **A latitudinal diversity gradient in planktonic marine bacteria**. *Proceedings of the National Academy of Sciences* 2008, **105**:7774-7778.

30. Bidle KD, Azam F: **Accelerated dissolution of diatom silica by marine bacterial assemblages**. *Nature* 1999, **397**:508-512.

31. Rooney-Varga JN, Giewat MW, Savin MC, Sood S, LeGresley M, Martin JL: **Links between phytoplankton and bacterial community dynamics in a coastal marine environment**. *Microb. Ecol* 2005, **49**:163-175.

32. West NJ, Obernosterer I, Zemb O, Lebaron P: **Major differences of bacterial diversity and activity inside and outside of a natural iron-fertilized phytoplankton bloom in the Southern Ocean**. *Environ Microbiol* 2008, **10**:738-756.

33. Lauro FM, McDougald D, Thomas T, Williams TJ, Egan S, Rice S, DeMaere MZ, Ting L, Ertan H, Johnson J, Ferriera S, Lapidus A, Anderson I, Kyrpides N, Munk AC, Detter C, Han CS, Brown MV, Robb FT, Kjelleberg S, Cavicchioli R: **The genomic basis of trophic strategy in marine bacteria**. *Proc. Natl. Acad. Sci. U.S.A* 2009, **106**:15527-15533.

34. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR, Bateman A: **The Pfam protein families database**. *Nucl. Acids Res.* 2010, **38**:D211-222.

35. Field D, Garrity 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, Glockner 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 S, 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: **The minimum information about a genome sequence (MIGS) specification**. *Nat Biotech* 2008, **26**:541-547.

36. Seshadri R, Kravitz SA, Smarr L, Gilna P, Frazier M: **CAMERA: A Community Resource for Metagenomics**. *PLoS Biology* 2007, **5**:e75.

37. Kottmann R, Kostadinov I, Duhaime MB, Buttigieg PL, Yilmaz P, Hankeln W, Waldmann J, Glöckner FO: **Megx.net: integrated database resource for marine ecological genomics**. *Nucleic Acids Res* 2010, **38**:D391-395.

38. R Development Core Team: *R: A Language and Environment for Statistical Computing*. Vienna, Austria: 2010.

39. Oksanen J, Blanchet FG, Kindt R, Legendre P, O'Hara RB, Simpson GL, Solymos P, Stevens MHH, Wagner H: *vegan: Community Ecology Package*. 2010.

40. Venables WN, Ripley BD: *Modern Applied Statistics with S*. Fourth. New York: Springer; 2002.

41. Rice P, Longden I, Bleasby A: **EMBOSS: the European Molecular Biology Open Software Suite**. *Trends Genet* 2000, **16**:276-277.

42. **Pfam: Home page** [http://pfam.sanger.ac.uk/].

43. Eddy SR: **A new generation of homology search tools based on probabilistic inference**. *Genome Inform* 2009, **23**:205-211.

44. Eddy SR: **HMMER3 beta test: User's guide**. 2009.

45. DeLong EF: **Microbial community genomics in the ocean**. *Nat Rev Micro* 2005, **3**:459-469.

46. Ramette A: **Multivariate analyses in microbial ecology**. *FEMS Microbiol. Ecol* 2007, **62**:142-160.

47. Minezaki Y, Homma K, Nishikawa K: **Genome-Wide Survey of Transcription Factors in Prokaryotes Reveals Many Bacteria-Specific Families Not Found in Archaea**. *DNA Res* 2005, **12**:269-280.

Figures

Figure 1 - Linear regression analysis of measured and interpolated environmental parameters.

Temperature (A), salinity (B) and salinity with sample GS033 removed (C). The points represent the samples. The solid blue line is the fitted linear function and the shaded area depicts the confidence interval for it.

Figure 2 - RDA biplot of TFs constrained by environment stability and space.

The ordination of TFs (in red) constrained by the explanatory variables (blue vectors) is shown. The lengths of the vectors correspond to the strength of the effect of that particular variable. RDA scaling 2 was used (scaling the TF scores). The angle between an explanatory variable vector and a TF (if a vector was to be drawn from the origin of the graph to this TF) approximates their correlation.

Tables

Table 1 - Multiple Regression results for single TFs.

Only results with correlation coefficient (multiple R-squared) above 0.3 (30% explained variation) are shown. The significance of each term in the linear model (p-value) is given next to it.

TF (non-DBD)	Multiple regression model	R-squared	p-value
response_reg	temperature (p<0.05) + phosphate (p<0.01) + X2^2 (p<0.05)	0.31	1.93E-03
peptidase_s24	temperature (p<0.001)	0.34	1.67E-02
pro_dh	temperature (p<0.001) + X1 (p<0.01)	0.38	6.23E-02
aldedh	temperature (p<0.001) + X2 (p<0.1)	0.46	3.78E-03
Sugar.bind	temperature (p<0.001) + salinity (p<0.01)	0.47	2.30E-03
utra	temperature (p<0.001)	0.49	1.06E-04
tobe	temperature (p<0.001) + salinity (p<0.05) + silicate (p<0.05) + X2 (p<0.05)	0.56	1.22E-03
lysr_substrate	temperature (p<0.001) + X2 (p<0.001)	0.58	1.63E-05
TF (DBD)	**Multiple regression model**	**R-squared**	**P-value**
laci	temperature (p<0.001) + silicate (p<0.05)	0.30	6.44E-04
laci	temperature (p<0.001) + phosphate (p<0.05)	0.31	4.88E-04

gntr	temperature (p<0.001) + X2 (p<0.05)	0.38	5.61E-02
penicillinase_r	temperature (p<0.05) + salinity (p<0.01) + phosphate (p<0.01) + X2 (p<0.05)	0.41	3.36E-04
hth_arac	temperature (p<0.001) + X2 (p<0.01)	0.41	1.87E-02
hth_6	temperature (p<0.001) + phosphate (p<0.05)	0.43	1.03E-05
hth_3	temperature (p<0.001) + silicate (p<0.1) + X2 (p<0.01)	0.52	1.42E-03
tetr_n	temperature (p<0.001) + X2 (p<0.01)	0.54	1.10E-04
trp_repressor	temperature (p<0.001) + phosphate (p<0.1)	0.55	8.52E-08
lyttr	temperature (p<0.01) + silicate (p<0.05) + X2 (p<0.01)	0.57	2.22E-04
hth_1	temperature (p<0.001) + salinity (p<0.05) + X2 (p<0.01)	0.60	5.65E-08

Additional files

Additional file 1 – Supplement_S1.pdf
A PDF file containing figures and tables that further describe and visualize the analysis in more detail.

Figures:

Figure S1.1: Distribution of SCGs against the number of sequences per sample.

Figure S1.2: Coefficient of variation of SCGs against the number of sequences per sample.

Figure S1.3: Seven descriptive statistic functions of SCG counts against the number of sequences per sample.

Figure S1.4: Correlation of environmental stability variables to each other.

Tables:

Table S1.1: A list of SCG models that were identified as outliers.

Table S1.2: Correlation coefficients of environmental stability variables

Table S1.3: A list of SCG HMMs based on Ciccarelli et al. (2006).

Table S1.4: TF models after Minezaki et al. (2005).

Additional file 2 – Supplement_S2.zip
A zip file containing data and R code for reproducing the analysis in this study. Contents are listed below:

Rcode.txt - R code used for the analysis in this publication
Supplement_S2.1.csv - Interpolated and measured values for temperature and salinity
Supplement_S2.2.csv - Monthly interpolations for GS041
Supplement_S2.3.csv - SCG raw counts
Supplement_S2.4.csv -Number of sequences per sample
Supplement_S2.5.csv - Environmental stability measures
Supplement_S2.6.csv - TF model categories (DBD. non-DBD)
Supplement_S2.7.csv - TF raw counts
Supplement_S2.8.csv - Cartesian distance between GOS samples

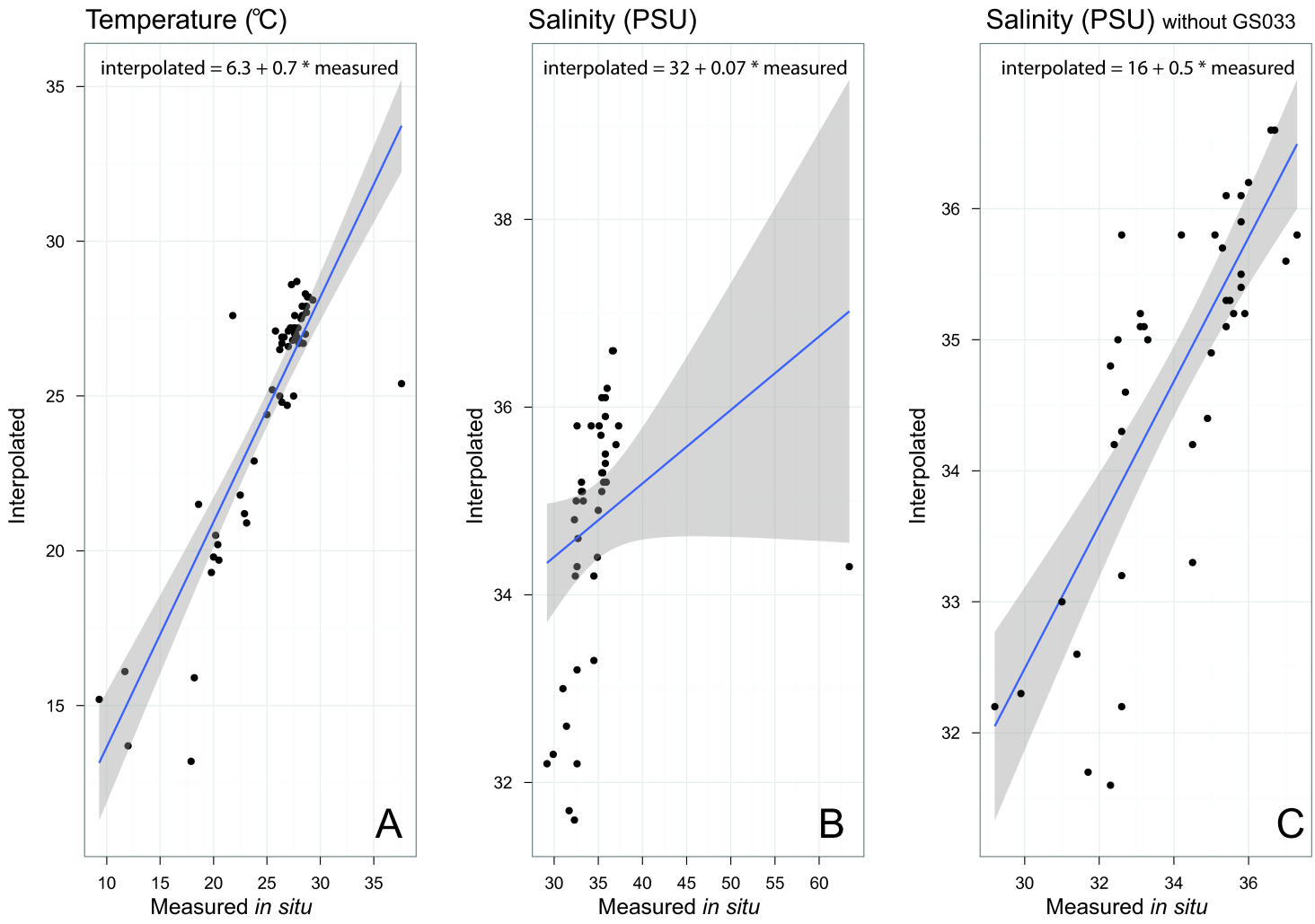

Temperature (℃)
interpolated = 6.3 + 0.7 * measured
Interpolated
Measured *in situ*
A
Salinity (PSU)
interpolated = 32 + 0.07 * measured
Interpolated
Measured *in situ*
B
Salinity (PSU) without GS033
interpolated = 16 + 0.5 * measured
Interpolated
Measured *in situ*
C

Figure 1

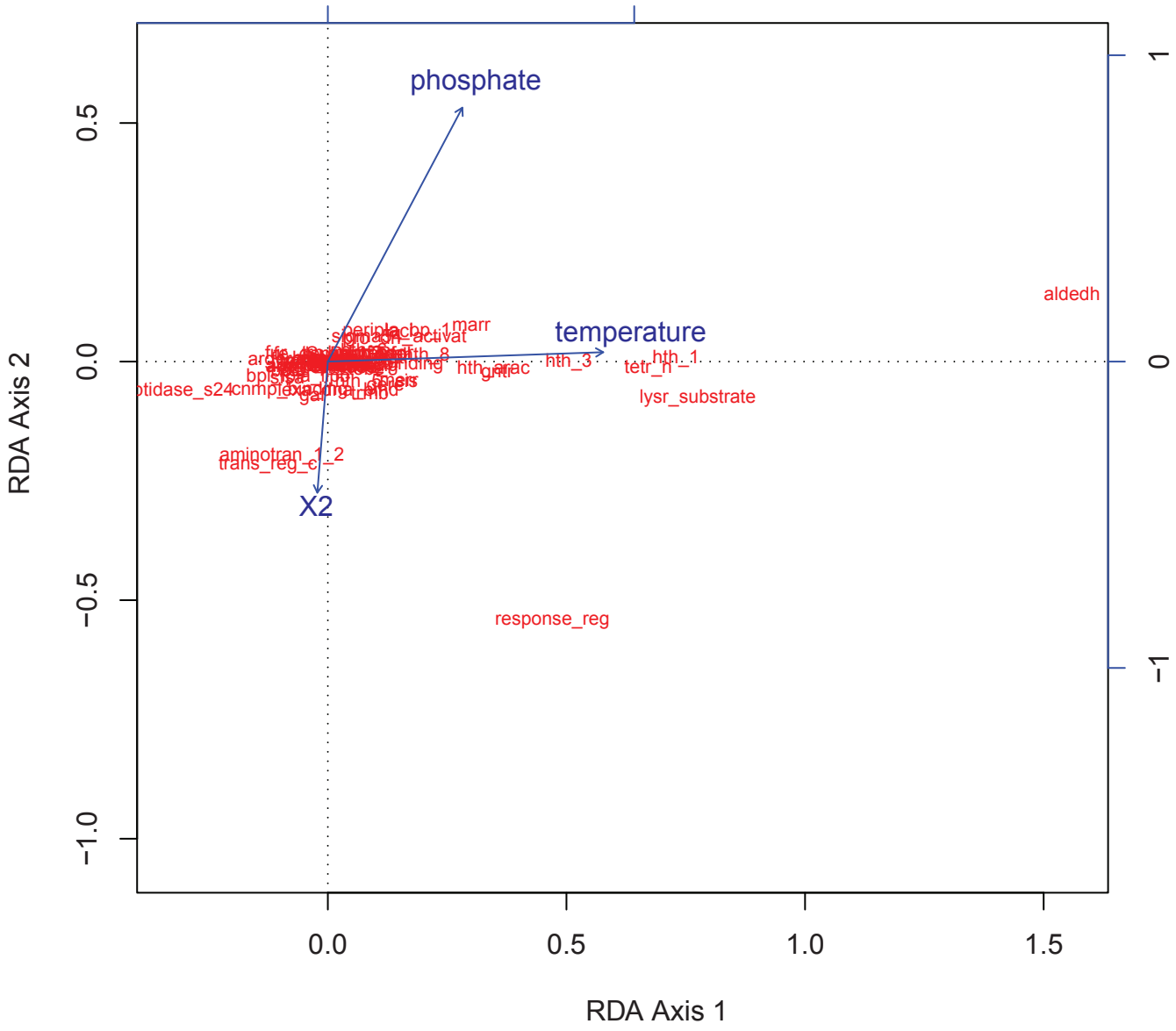

phosphate
temperature
X2
aldedh
lysr_substrate
response_reg
aminotran_1_2
trans_reg_c
hth_3
tetr_fl
hth_1
marr
RDA Axis 2
RDA Axis 1
0.5
0.0
−0.5
−1.0
0.0
0.5
1.0
1.5
1
0
−1

Figure 2

Supplement S1

List of Figures:

List of Tables:

Supplement S1

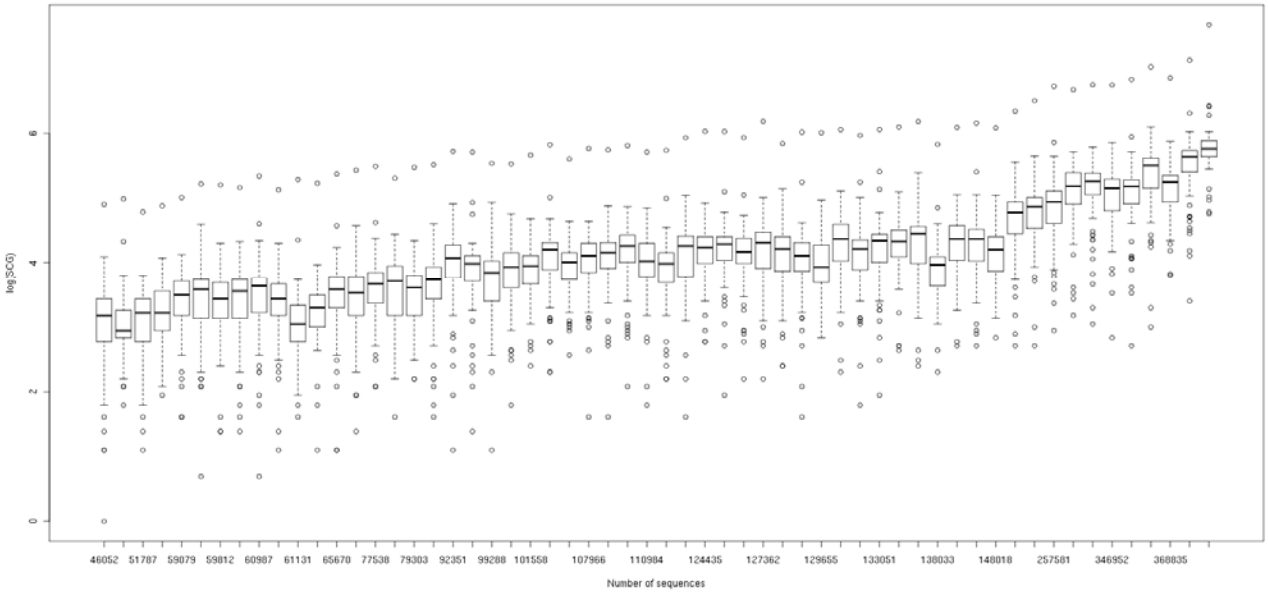


Figure S1.1: Distribution of SCGs against the number of sequences per sample.
The absolute counts of SCGs per sample were log-transformed (Y axis). In the boxplots, the whiskers' ends correspond roughly to ±2 standard deviations around the mean. More concretely, they denote the furthest data points still within 1.5 times the interquartile range (IQR) of the first (Q1) and the third quartile (Q3). The IQR is calculated as follows: IQR = Q3-Q1. The dots represent SCGs that lie outside these ranges and are therefore considered outliers. A list of the outliers is available in Supplement Table S1.1.

Supplement S1

Table S1.1: A list of SCG HMM models that were identified as outliers.
The number of samples in which the model was an outlier and the percentage of all samples (58 in total) are presented.

SCG model	number of samples	percent of all samples
Above 1.5 IQR of Q3		
usg	1	2
if_n2	4	7
reca	18	31
ruvb_n	58	98
Below 1.5 IQR of Q1		
rimm	1	2
secg	4	7
ruvc	6	10
duf150	7	12
trigger_c	9	15
exonuc_vii_s	15	25
tyr_deacylase	20	34
glutr_n	31	53
duf177	35	59
hrca	41	69
glutr_dimer	46	78
ribosomal_s20p	55	93

Supplement S1

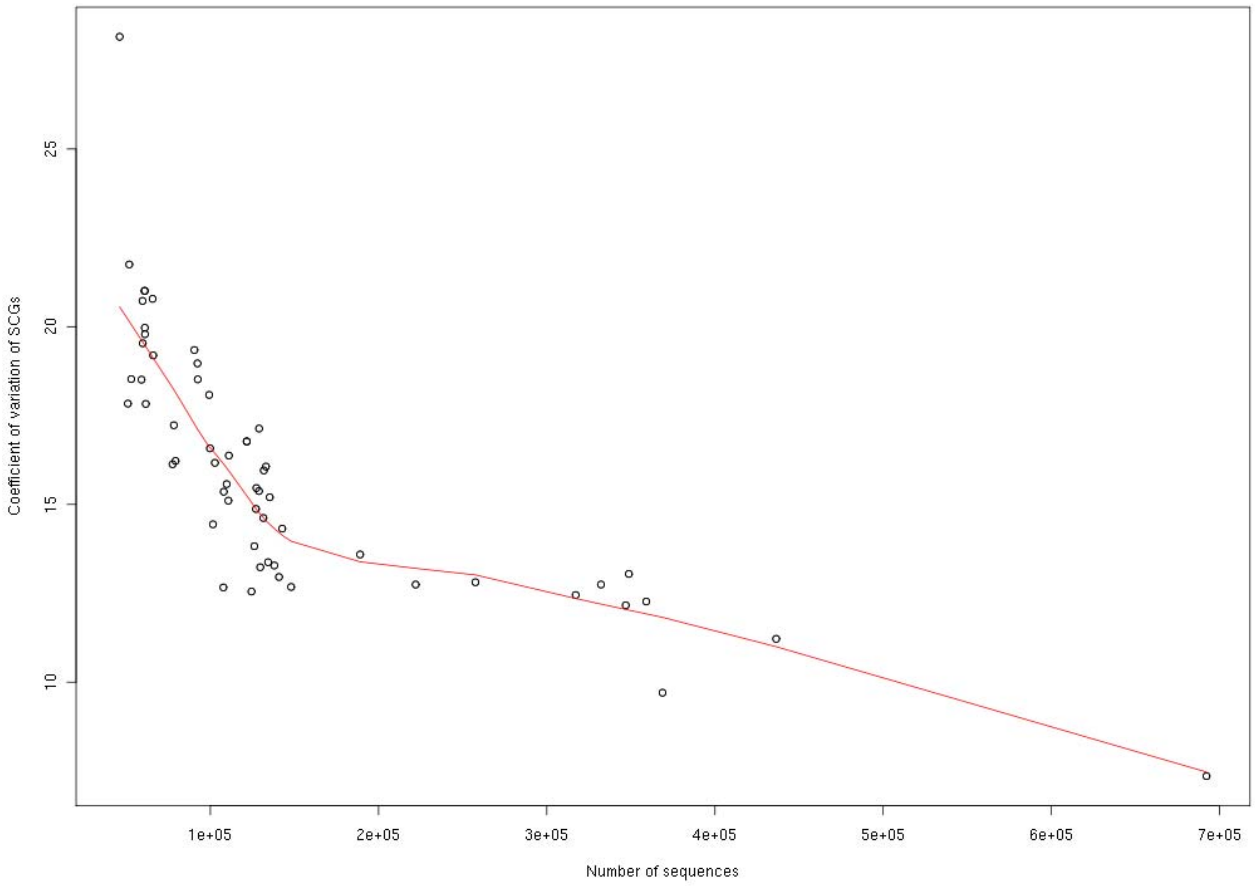


Figure S1.2: Coefficient of variation of SCGs against the number of sequences per sample. The variation within SCG numbers decreases with increasing number of sequences, supporting the idea that deeper sequencing delivers more stable data.

Supplement S1

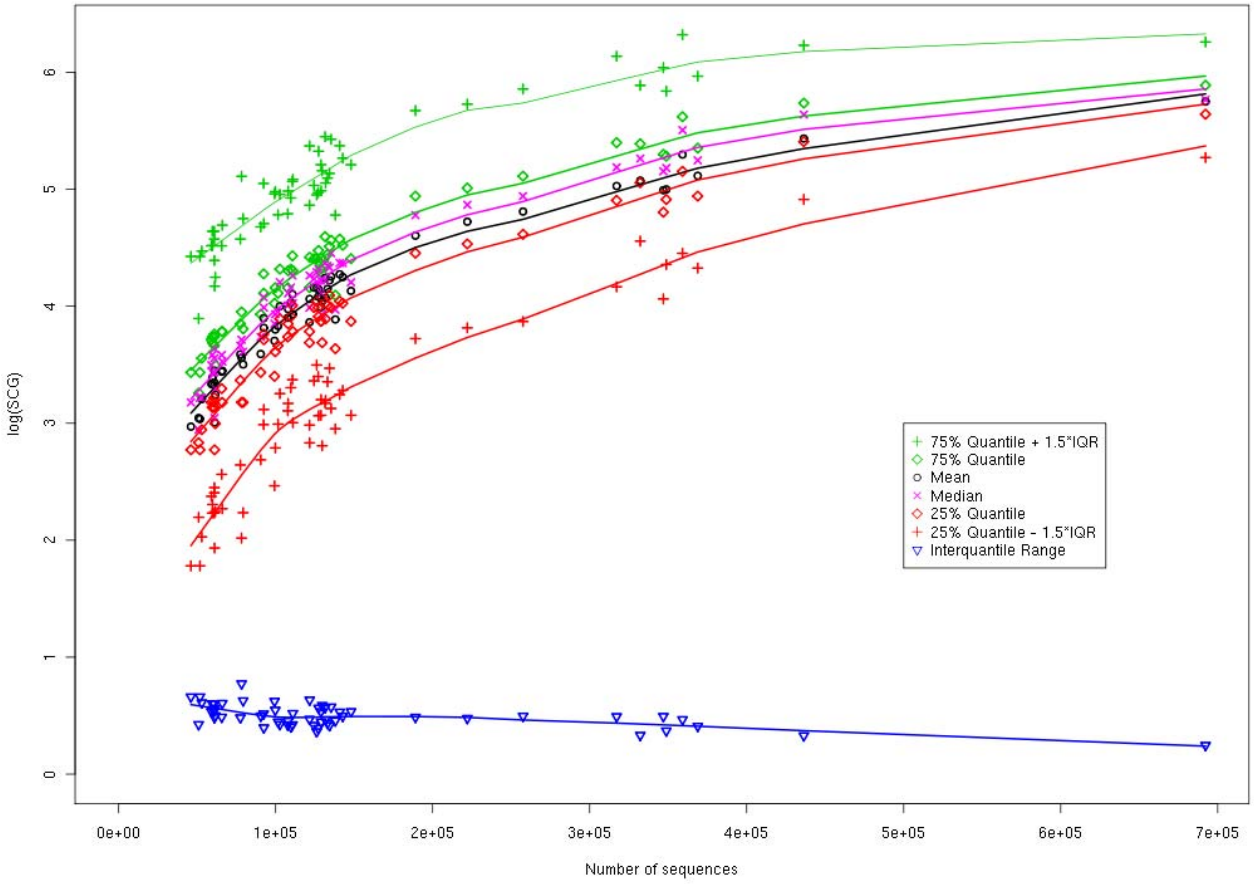


Figure S1.3: Seven descriptive statistic functions of SCG counts against the number of sequences per sample.
The absolute counts of SCGs per sample were log-transformed (Y axis).

Supplement S1

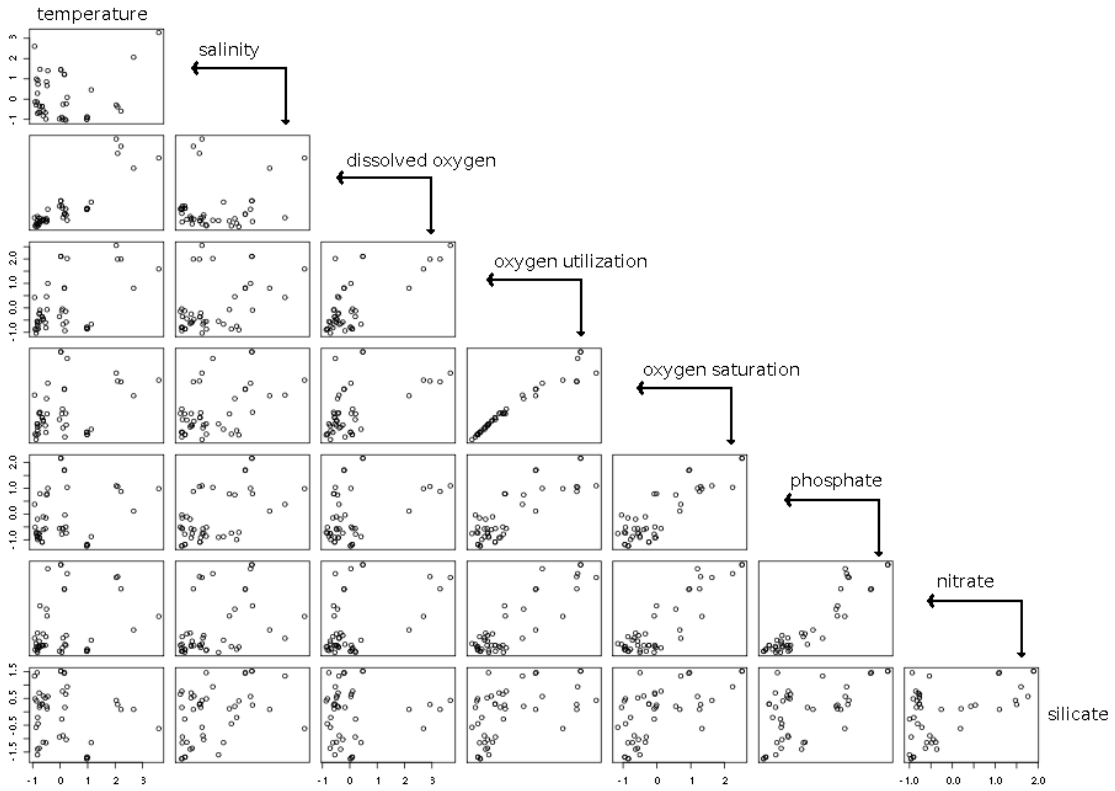


Figure S1.4 Correlation of environmental stability variables.
This is a visual representation - a roughly diagonal line in any direction would mean a considerable correlation.

Table S1.2 Correlation coefficients of environmental stability variables.
Variable pairs with Spearman correlation coefficient above 0.6 are shown.

stability measures		rho	p-value
temperature	oxygen_dissolved	0.81	3.27E-011
oxygen_utilization	oxygen_saturation	0.99	2.20E-016
oxygen_utilization	phosphate	0.70	1.31E-007
oxygen_saturation	phosphate	0.70	1.31E-007
oxygen_utilization	nitrate	0.66	9.51E-007
oxygen_saturation	nitrate	0.66	8.97E-007
phosphate	nitrate	0.85	2.96E-013

Supplement S1

Table S1.3: A list of SCG HMMs based on Ciccarelli et al. (2006).

Accession	Pfam Id	Model length	Average domain length
PF00189	Ribosomal_S3_C	85	82.0
PF00252	Ribosomal_L16	133	113.7
PF00417	Ribosomal_S3_N	66	63.3
PF00453	Ribosomal_L20	108	101.7
PF00475	IGPD	145	144.3
PF00542	Ribosomal_L12	68	67.0
PF00584	SecE	57	56.6
PF00745	GlutR_dimer	101	100.6
PF00825	Ribonuclease_P	111	109.0
PF00829	Ribosomal_L21p	96	95.0
PF00831	Ribosomal_L29	58	57.4
PF00886	Ribosomal_S16	62	57.7
PF00889	EF_TS	221	180.3
PF01016	Ribosomal_L27	81	80.7
PF01192	RNA_pol_Rpb6	57	54.4
PF01196	Ribosomal_L17	97	100.8
PF01245	Ribosomal_L19	113	113.0
PF01250	Ribosomal_S6	92	91.7
PF01281	Ribosomal_L9_N	48	47.9
PF00828	Ribosomal_L18e	129	118.9
PF01628	HrcA	224	219.4
PF01649	Ribosomal_S20p	84	82.0
PF01668	SmpB	68	67.2
PF01746	tRNA_m1G_MT	186	190.6
PF01765	RRF	165	163.1
PF01782	RimM	84	83.7
PF02033	RBFA	104	104.9
PF02075	RuvC	149	147.7
PF02092	tRNA_synt_2f	549	541.7
PF02130	UPF0054	145	142.1
PF02132	RecR	41	41.0
PF02357	NusG	92	98.2
PF02410	DUF143	100	98.5
PF02542	YgbB	157	156.4
PF02565	RecO_C	118	151.4
PF02576	DUF150	141	138.8
PF02580	Tyr_Deacylase	145	142.8
PF02609	Exonuc_VII_S	53	52.9
PF02620	DUF177	119	114.4
PF02686	Glu-tRNAGln	72	72.4
PF02912	Phe_tRNA-synt_N	73	72.6
PF02978	SRP_SPB	104	100.3
PF03147	FDX-ACB	94	94.2
PF03483	B3_4	174	167.7
PF03484	B5	70	70.0
PF03726	PNPase	83	81.9
PF03840	SecG	74	73.3
PF03948	Ribosomal_L9_C	87	86.9
PF04760	IF2_N	54	52.0

Supplement S1

PF05201	GlutR_N	152	148.4
PF05496	RuvB_N	234	212.7
PF05698	Trigger_C	162	154.7
PF00154	RecA	323	233.9

Supplement S1

Table S1.4: A list of TF HMMs based on Minezaki et al. (2005).

Accession	Pfam Id
PF00027	cNMP_binding
PF00072	Response_reg
PF00126	HTH_1
PF00155	Aminotran_1_2
PF00158	Sigma54_activat
PF00165	HTH_AraC
PF00171	Aldedh
PF00196	GerE
PF00325	Crp
PF00356	LacI
PF00376	MerR
PF00392	GntR
PF00440	TetR_N
PF00480	ROK
PF00486	Trans_reg_C
PF00532	Peripla_BP_1
PF00717	Peptidase_S24
PF01022	HTH_5
PF01047	MarR
PF01316	Arg_repressor
PF01325	Fe_dep_repress
PF01340	MetJ
PF01371	Trp_repressor
PF01380	SIS
PF01381	HTH_3
PF01402	RHH_1
PF01418	HTH_6
PF01475	FUR
PF01590	GAF
PF01619	Pro_dh
PF01722	BolA
PF01726	LexA_DNA_bind
PF07702	UTRA

Accession	Pfam Id
PF01965	DJ-1_PfpI
PF01978	TrmB
PF02082	Rrf2
PF02237	BPL_C
PF02311	AraC_binding
PF02742	Fe_dep_repr_C
PF02805	Ada_Zn_binding
PF02863	Arg_repressor_C
PF02954	HTH_8
PF03099	BPL_LplA_LipB
PF03459	TOBE
PF03466	LysR_substrate
PF03472	Autoind_bind
PF03551	PadR
PF03704	BTAD
PF03749	SfsA
PF03965	Pencillinase_R
PF04023	FeoA
PF04198	Sugar-bind
PF04299	FMN_bind_2
PF04397	LytTR
PF04967	HTH_10
PF05068	MtlR
PF05247	FlhD
PF05443	ROS_MUCR
PF05848	CtsR
PF06018	CodY
PF06338	ComK
PF06506	PrpR_N
PF06923	GutM
PF06956	RtcR
PF06988	NifT
PF07417	Crl

2.4 Ecological perspectives on domains of unknown function: a marine point of view

Authors: Pier Luigi Buttigieg, Wolfgang Hankeln, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Melissa Beth Duhaime, and Frank Oliver Glöckner

Submitted to: The ISME Jorunal

Contribution: HMM calculations

Relevance: Generating hypothesis about the possible functions of protein domains based on their co-occurrence patterns and environmental gradients.

Ecological perspectives on domains of unknown function: a marine point of view

Pier Luigi Buttigieg,[1,2] Wolfgang Hankeln,[1,2] Ivaylo Kostadinov,[1,2] Renzo Kottmann,[1] Pelin Yilmaz,[1,2] Melissa Beth Duhaime,[1,2] and Frank Oliver Glöckner[1,2]

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

[2] Jacobs University Bremen gGmbH, D-28759, Bremen, Germany

Correspondence should be addressed to PLB (pbuttigi@mpi-bremen.de)

Metagenomic datasets from environmental samples offer attractive opportunities to characterize genomic elements of unknown function. We employed graph-theoretic approaches to visualize correlations between protein domains of unknown function detected in the Global Ocean Sampling metagenomes. Functional hypotheses for groups of these domains were generated based on network topology and existing putative functional assignments. Environmental contextualization of one such hypothesis was carried out using indirect gradient analysis.

Genomic and metagenomic sequencing projects are revealing ever-increasing numbers of novel genes, many of unknown function. The Pfam 23 database (Finn et al, 2008), for example, stored some 10 340 protein domain families derived from conserved sequence data with 22% dubbed “domains of unknown function” (DUFs). This proportion is predicted to soon overtake that of functionally characterized domains (Bateman et al, 2010), prompting calls for community action (Roberts, 2004). In their response, Jaroszewski *et al.* (2009) and Goonesekere *et al.* (2010) noted several DUFs that appeared to be variations of functionally characterized protein folds, most likely maintained due to an extension of an organism’s ecological niche. It is reasonable to expect that conserved DUFs enhance ecological performance; however, characterizing DUFs from an ecological perspective has yet to be attempted. In this communication, we present a method of functional attribution based on DUF correlation across the Global Ocean Sampling (GOS) metagenome collection (Rusch et al, 2007). Network visualizations were used in hypothesis generation followed by indirect gradient analysis to contextualize one hypothesis with environmental metadata. Together, these approaches aim to support efforts in DUF characterization using ecogenomic resources.

Correlation analysis of microbial taxa and environmental parameters has previously been used to construct association networks (Fuhrman & Steele, 2008; Fuhrman, 2009). Just as the correlation of taxa-abundance may elucidate a given taxon's ecosystem-level interactions and function, correlation of protein domains across environments may grant insight into their potential associations and roles. This approach parallels the identification of unknown metabolic modules whereby genomic features found to co-vary in response to experimental perturbations are grouped in putative metabolic modules (Breitling et al, 2008). To detect such associations in metagenomic datasets, we measured the Spearman rank correlation (ρ) between DUFs detected in the globally-distributed GOS metagenomes (473 351 DUFs detected in 454 varieties across 79 metagenomes, see **Supplementary methods** online). We visualized these results as network graphs. Vertices (representing DUFs varieties) were connected if their ρ was ≥ 0.90. As abundances of 454 DUF varieties were correlated, we enforced a Bonferroni-corrected p-value cut-off of $\sim 2.20 \times 10^{-5}$ (0.01 / 454). We embedded the graph using the Fruchterman-Reingold procedure (Fruchterman & Reingold, 1991). A minimal spanning tree was visualized after Prim's algorithm (Prim, 1957) to aid visual interpretation (**Fig. 1b**). We assigned DUFs to putative functional categories guided by Pfam descriptions and linked literature, color-coding vertices accordingly.

We observed two prominent networks, one dominated by DUFs linked to photosynthetic organisms (**Fig. 1, II**) and another comprised of more diverse members (**Fig. 1, I**). Smaller networks were observed, including one associating DUFs 403, 404 and 407 (**Fig. 1, III**), domains known to co-occur[4]. Employing a 'guilty by association' approach (Merico et al, 2009), we propagated hypotheses across closely-embedded domains. We thus hypothesized that DUFs in network II (**Fig. 1**), including unassigned DUFs, describe a microbial photoreactivity module. The larger network (**Fig. 1, I**) presented difficulty in interpretation due to its members' diverse functions; however, it suggests a functional constellation suited to the marine epipelagic ecosystem. DUFs with putative functions in Respiration, Cell Division and Cell Cycle, Regulation and Cell signaling, and Membrane Transport are present, with the EamA domain (PF00892, formerly 'DUF6') at its centre. Although still uncharacterized, EamA occurs in the plant pathogen *Erwinia chrysanthemi*'s PecM protein, involved in pectinase, cellulase and pigment regulation (Jack et al, 2001). Such regulation is well-suited to varying nutrient availability, primary productivity, and irradiation in the water column.

Hypotheses generated from covariation across metagenomes may be contextualized with environmental data to enhance interpretation. We employed indirect gradient analysis

after Virtanen *et al.*(2006) to relate DUF abundances in network II (**Fig. 1;** 17 DUF varieties) to chlorophyll concentrations at appropriate GOS sites (n=56; see **Supplementary methods** online). We standardized DUF abundances at each site by the median abundance of 22 'single-copy domains' detected at that site. We then ordinated sites by non-metric dimensional scaling (NMDS; **Fig. 2**, hollow circles) using Bray-Curtis dissimilarities. Next, we performed a least squares, linear fit of chlorophyll data with significance (P(>R)) determined by permutation (n=1000). To explore non-linear relationships between chlorophyll concentrations and the ordination, we visualized generalized additive model (GAM) fits as smoothed, non-parametric isoclines (**Fig. 2**) with significance determined by ANOVA (Wood, 2008). After Virtanen *et al.*, we interpreted coefficients of determination (R^2) as goodness-of-fit measures for linear vectors (R_v^2) and non-parametric surfaces (R_s^2). Analyses were performed in R (http://www.r-project.org). We observed that these DUF abundances moderately, but significantly, structure GOS sites along chlorophyll concentration ($R_v^2 \approx 0.52$, $P(>R) \approx 9.99 \times 10^{-4}$; $R_s^2 \approx 0.91$, $p \approx 2.00 \times 10^{-16}$). An improved, albeit less significant, fit ($R_v^2 \approx 0.64$, $P(>R) \approx 4.00 \times 10^{-3}$; $R_s^2 \approx 0.98$, $p \approx 5.8 \times 10^{-2}$) and a more even resolution of sites may be observed when ordinating geographically localized sample groups such as that along the North American East Coast (GS002, GS004-8, GS012-14; n=9, plot not shown). The GAM surface reveals considerable non-linear effects below chlorophyll concentrations of ~2.0 µg kg^{-1} seawater, where most sites – particularly from oligotrophic waters – are ordinated. Such effects may rise from the diverse functions, multi-functionality, and the selective interactions between elements in biological systems (Kitano, 2002). The global coverage of GOS, across numerous ecoregions, may also introduce unexpected variation. Nonetheless, if these chlorophyll measurements are understood as a proxy for phytoplankton abundance, our results tentatively support hypotheses linking the functional community structure described by these DUFs to the abundance of photoreactive plankton. This manner of environmental contextualization may provide useful perspectives on the function of microbial genomic features in their surrounding ecosystems.

Ecogenomic datasets promise to deliver valuable insight into the roles of uncharacterized genes and proteins and await the application of exploratory meta-analysis. The prospects are greater if future 'omics' sampling is performed along clear environmental gradients and accompanied by comprehensive and standardized metadata (Field et al, 2008). Here, we demonstrated the use of graph theory and numerical ecology to offer new, contextualized hypotheses on uncharacterized targets for community investigation. Caution in interpretation, alongside strict detection criteria, is encouraged to reduce association fallacies

and hasty generalizations. Subsequently, wet-lab validation and falsification of these *in silico* results are required to establish standards for future predictions. Navigating the topology of ecogenomic space will be challenging; however, its immense potential in guiding biological enquiry warrants interdisciplinary attention.

Supplementary information is available at the ISME Journal's website (http://www.nature.com/ismej)

Bateman A, Coggill PC, Finn RD. (2010). DUFs: families in search of function. *Acta Crystallogr Sect F Struct Biol Cryst Commun* **66**:1148-1152.

Breitling R, Vitkup D, Barrett MP. (2008). New surveyor tools for charting microbial metabolic maps. *Nat Rev Microbiol* **6**:156-161.

Field D, Garrity G, Gray T, Morrison N, Selengut J, Sterk P et al. (2008). The minimum information about a genome sequence (MIGS) specification. *Nat Biotechnol* **26**:541-547.

Finn RD, Tate J, Mistry J, Coggill PC, Sammut SJ, Hotz H et al. (2008). The Pfam protein families database. *Nucleic Acids Res* **36**:D281–D288.

Fuhrman JA, Steele J. (2008). Community structure of marine bacterioplankton: patterns, networks, and relationships to function. *Aquat Microb Ecol* **53**:69-81.

Fuhrman JA. (2009). Microbial community structure and its functional implications. *Nature* **459**:193-199.

Fruchterman TMJ, Reingold EM. (1991). Graph drawing by force-directed placement. *Software Pract Ex* **21**:1129-1164.

Goonesekere NCW, Shipely K, O'Connor K. (2010). The challenge of annotating protein sequences: The tale of eight domains of unknown function in Pfam. *Comput Biol Chem*. **34**:210-214.

Jack DL, Yang NM, Saier MH. (2001). The drug/metabolite transporter superfamily. *Eur J Biochem* **268**:3620-3639.

Jaroszewski L, Li Z, Krishna SS, Bakolitsa C, Wooley J, Deacon AM et al. (2009). Exploration of uncharted regions of the protein universe. *PLoS Biol* **7**:e1000205.

Kitano H. (2002). Computational systems biology. *Nature* **420**:206-210.

Merico D, Gfeller D, Bader GD. (2009). How to visually interpret biological data using networks. *Nat Biotechnol* **27**:921-9244.

Prim RC. (1957). Shortest connection networks and some generalizations. *The Bell System Technical Journal* **36**:1389-1401.

Roberts RJ. (2004). Identifying protein function - a call for community action. *PLoS Biol*. **2**:e42.

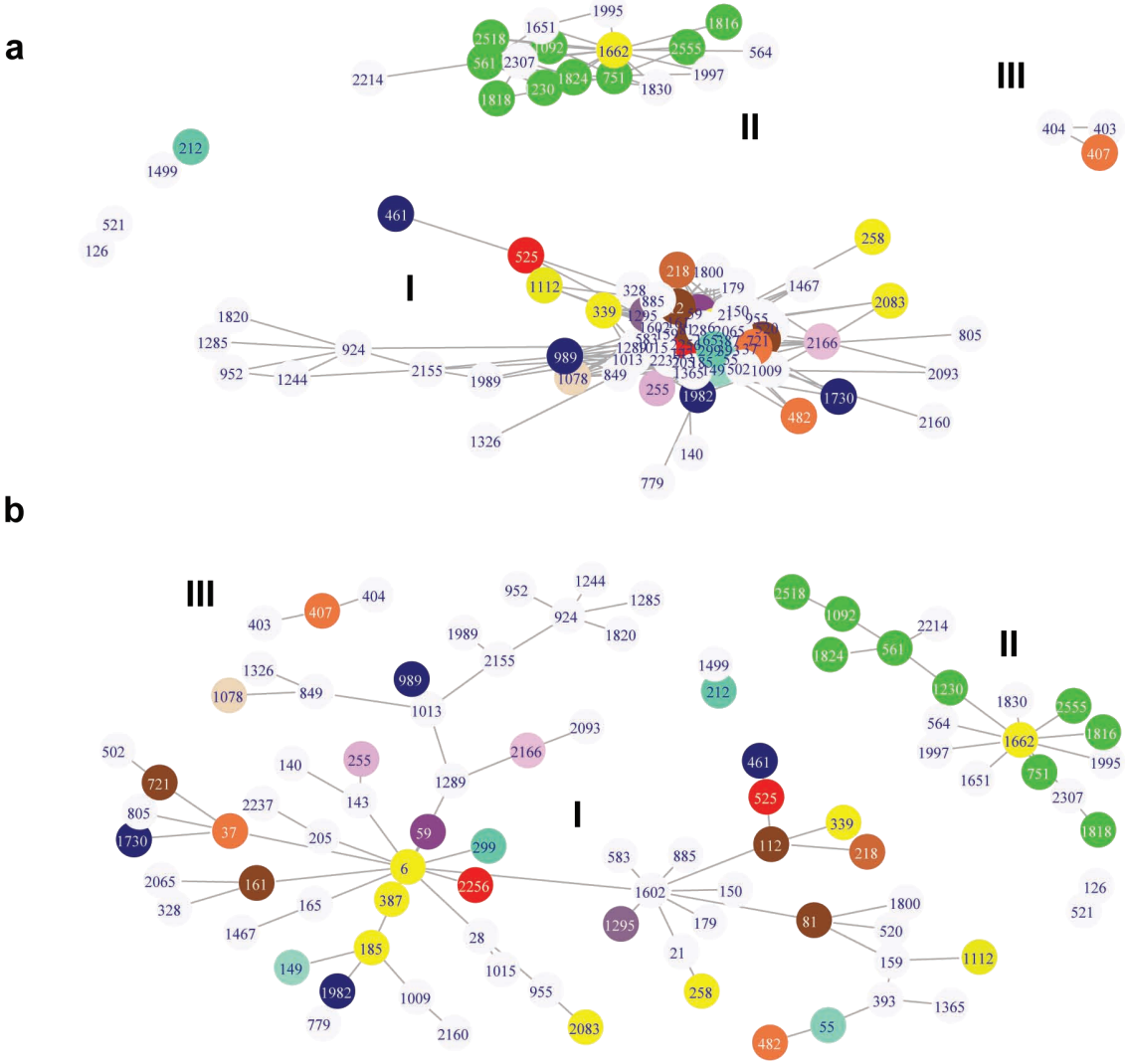
Rusch DB, Halpern AL, Sutton G, Heidelberg KB, Williamson S, Yooseph S et al. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biol* **5**:e77.

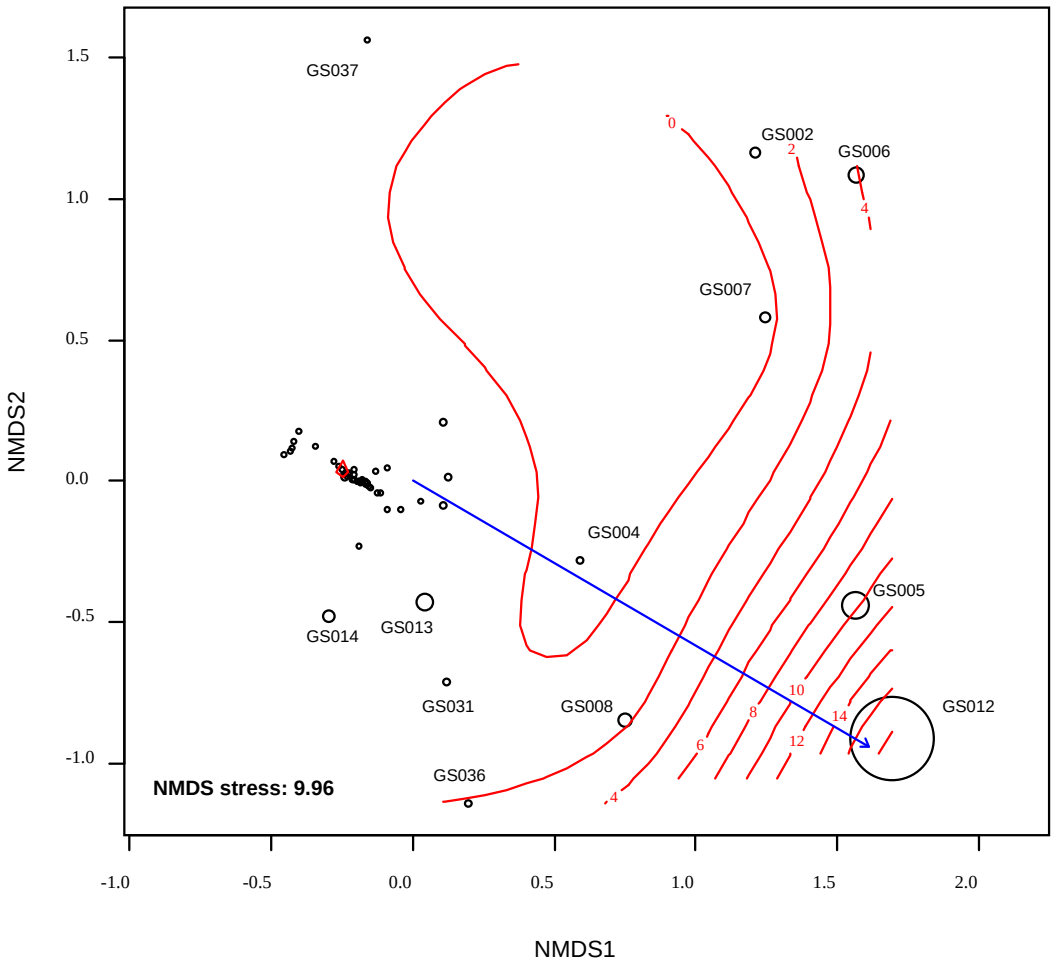
Virtanen R, Oksanen J, Oksanen L, Razzhivin VY. (2006). Broad-scale vegetation-environment relationships in Eurasian high-latitude areas. *J Veg Sci* **17**:519-528.

Wood SN. (2008). Fast stable direct fitting and smoothness selection for generalized additive models. *J R Stat Soc Series B Stat Methodol* **70**:495-518.

Figure 1: Hypothesis generation using network representations of Spearman rank correlations between DUF abundances across GOS metagenomes. Vertices are labelled with the corresponding DUF number (i.e. "59" represents DUF59). Network I includes DUFs with a variety of functions, possibly involved in a response to nutrient input in the marine epipelagic zone. Network II is dominated by DUFs linked to photosynthetic organisms and functions. Network III is composed of three DUFs known to co-occur. a) Fruchterman-Reingold embedded network. Edge lengths are inversely related to correlation strength b) Minimum spanning tree representation of DUF correlations. Only edges describing the shortest path (hence, strongest correlation) between vertices are visualized.

Figure 2: Hypothesis contextualization using indirect gradient analysis of GOS sites described by the abundances of DUFs in Network II (ref. Fig. 1) and chlorophyll concentration data. Each bubble represents one GOS site and bubble size reflects the *in situ* chlorophyll concentration measured during the expedition. Bubble positions reflect the Bray-Curtis dissimilarity between sites calculated from the per site abundance profiles of DUFs in Network II. The blue vector describes the linear fit of chlorophyll concentration data to the ordination ($R_v^2 \approx 0.52$, $P(>R) \approx 9.99 \times 10^{-4}$). Red isoclines describe a generalized additive model fit of the same chlorophyll data to the ordination ($R_s^2 \approx 0.91$, $p \approx 2.00 \times 10^{-16}$) and are labeled with the corresponding chlorophyll concentration (in μg chlorophyll per kg seawater). Regions where the chlorophyll isoclines and vector intersect perpendicularly suggest a coherent response gradient of metagenomic DUF content to chlorophyll concentrations.


a
I
II
III
b
I
II
III
Unknown
Photosynthesis
Respiration
Regulation and Cell signalling
Cell Division and Cell Cycle
DNA Metabolism
RNA Metabolism
Phosphorus Metabolism
Amino Acids and Derivatives
Protein Metabolism
Fatty Acids Lipids and Isoprenoids
Metabolism of Aromatic Compounds
Cofactors Vitamins Prosthetic Groups Pigments
Secondary Metabolism
Cell Wall and Capsule
Membrane Transport
Motility and Chemotaxis
Stress Response


GS037
GS002
GS006
GS007
GS004
GS005
GS014
GS013
GS031
GS008
GS012
GS036
0
2
4
6
8
10
12
14
NMDS stress: 9.96
NMDS1
NMDS2
1.5
1.0
0.5
0.0
-0.5
-1.0
-1.0
-0.5
0.0
0.5
1.0
1.5
2.0

2.5 Phylogenetic Characterisation of Picoplanktonic Populations with High and Low Nucleic Acid Content in the North Atlantic Ocean

Authors: Martha Schattenhofer, Jörg Wulf, Ivalyo Kostadinov, Frank Oliver Glöckner, Mikhail V. Zubkov, Bernhard M. Fuchs
Published in: Systematic and Applied Microbiology, 2011 (in press)
Contribution: genomic data collection and integration
Relevance: A classic molecular ecology study of bacterial phytoplankton. Integrated metadata from megx.net was used.

Elsevier Editorial System(tm) for Systematic and Applied Microbiology
Manuscript Draft

Manuscript Number: SAM 3399R1

Title: Phylogenetic Characterisation of Picoplanktonic Populations with High and Low Nucleic Acid Content in the North Atlantic Ocean

Article Type: Full Length Papers

Section/Category: Applied and Ecological Microbiology

Keywords: Flow cytometry; CARD-FISH; prokaryotic picoplankton; marine; phylogenetic composition; genome size

Corresponding Author: Dr. Bernhard Maximilian Fuchs,

Corresponding Author's Institution: Max Planck Institute for Marine Microbiology

First Author: Martha Schattenhofer, Dr.

Order of Authors: Martha Schattenhofer, Dr.; Jörg Wulf; Ivalyo Kostadinov; Frank Oliver Glöckner, Prof.; Mikhail V Zubkov, Dr.; Bernhard Maximilian Fuchs

Abstract: In flow cytometric analyses of marine prokaryotic picoplankton often two populations with distinct differences in their apparent nucleic acid content are discernable, one with a high and one with a low nucleic acid content (HNA and LNA, respectively). In this study we determined the phylogenetic composition of flow cytometrically sorted HNA and LNA populations, collected at six stations along a transect across three oceanic provinces from Iceland to the Azores. Catalyzed reporter deposition fluorescence in situ hybridization (CARD-FISH) analysis of sorted cells revealed distinct differences in phylogenetic composition between the LNA and HNA populations with only little overlap. At all stations the LNA population was dominated by the alphaproteobacterial clade SAR11 (45 - 74%). Also, Betaproteobacteria were always present at 2-4%. While the LNA composition was rather stable, the HNA populations were composed of distinct phylogenetic clades in the different oceanic provinces of Arctic and Tropics. For example Cyanobacteria dominated the North Atlantic Gyre HNA population (29 - 44%) with Prochlorococcus as the major clade (34 - 44%), but were low in Arctic and Polar waters (1% and 5%, respectively). In contrast, Bacteroidetes accounted for the majority of HNA cells in the Polar and Arctic province (26% and 32%, respectively), but were low in the Gyre region (3 - 10%). The DNA content of the HNA population was about 3.5 times higher than that of the LNA populations. This reflects differences in the genome sizes of closely related cultured representatives of HNA clades (3-6 Mbp) and LNA clades (1.3-1.5 Mbp).

Phylogenetic Characterisation of Picoplanktonic Populations with High and Low Nucleic Acid Content in the North Atlantic Ocean

Martha Schattenhofer[1+], Jörg Wulf[1], Ivalyo Kostadinov[2,4], Frank Oliver Glöckner[2,4], Mikhail V. Zubkov[3], Bernhard M. Fuchs[1*]

[1] Department of Molecular Ecology, and

[2] Microbial Genomics Group, Max Planck Institute for Marine Microbiology, Bremen, Germany

[3] National Oceanography Centre, Southampton, United Kingdom

[4] Jacobs University Bremen, Bremen, Germany

Running head: Phylogenetic characterisation of cytometric populations

+ present address:
Department of Environmental Microbiology, UFZ - Helmholtz Centre for Environmental Research, Leipzig, Germany

* Corresponding author:
Address: Celsiusstr. 1, D-28359 Bremen, Germany
Email address: bfuchs@mpi-bremen.de
Telephone: +49 421 2028 935
Fax: +49 421 2028 790

Abstract

In flow cytometric analyses of marine prokaryotic picoplankton often two populations with distinct differences in their apparent nucleic acid content are discernable, one with a high and one with a low nucleic acid content (HNA and LNA, respectively). In this study we determined the phylogenetic composition of flow cytometrically sorted HNA and LNA populations, collected at six stations along a transect across three oceanic provinces from Iceland to the Azores. Catalyzed reporter deposition fluorescence in situ hybridization (CARD-FISH) analysis of sorted cells revealed distinct differences in phylogenetic composition between the LNA and HNA populations with only little overlap. At all stations the LNA population was dominated by the alphaproteobacterial clade SAR11 (45 – 74%). Also, Betaproteobacteria were always present at 2-4%. While the LNA composition was rather stable, the HNA populations were composed of distinct phylogenetic clades in the different oceanic provinces of Arctic and Tropics. For example Cyanobacteria dominated the North Atlantic Gyre HNA population (29 – 44%) with Prochlorococcus as the major clade (34 – 44%), but were low in Arctic and Polar waters (1% and 5%, respectively). In contrast, Bacteroidetes accounted for the majority of HNA cells in the Polar and Arctic province (26% and 32%, respectively), but were low in the Gyre region (3 – 10%). The DNA content of the HNA population was about 3.5 times higher than that of the LNA populations. This reflects differences in the genome sizes of closely related cultured representatives of HNA clades (3-6 Mbp) and LNA clades (1.3-1.5 Mbp).

Key words: Flow cytometry; CARD-FISH; prokaryotic picoplankton; marine; phylogenetic composition; genome size

Introduction

In the last two decades flow cytometry (FCM) has become a routine tool for reliable analysis and enumeration of picoplanktonic cells. The cells present in a water sample are often characterized by their scatter properties, the so called forward and side scatter, and their DNA content. General DNA stains are bright enough to yield high signal to noise ratios which enable the analysis of microbial cells and even viruses by flow cytometry [2, 18]. Since the first analysis of marine prokaryotic picoplankton, a conspicuous recurring pattern was visible in a scatter versus DNA fluorescence dotplot diagram. Two populations were discernable, one with a high and one with a low DNA content. The terms LDNA or LNA for 'low nucleic acid content bacteria' and HDNA or HNA for 'high nucleic acid content bacteria' have been coined for these cytometrically defined populations [7, 14, 16, 25]. However, the biological nature and ecological function of such populations remained poorly understood. While the HNA cells have been thought to represent the active part of the microbial community the role of the LNA cells has been controversial. Some researchers reported them either as inactive, dead or fragmented cells [14, 15, 23], but others showed that LNA cells can be as viable and active as HNA cells [33]. Several scenarios of how the LNA and HNA fractions are related to each other were proposed: (i) cells switch from one phenotype to the other i.e. start growing and develop from LNA to HNA cells, (ii) LNA cells represent dormant variants of HNA cells or (iii) LNA and HNA cells represent phylogenetically distinct groups of microorganisms with a fixed DNA content or (iv) a mixture of all of the three scenarios [1].

In previous studies phylogenetic analyses of the HNA and LNA populations have shown that both, the LNA and HNA populations are mainly composed of the same clades [23, 24] favouring scenario (i) or (ii). Our own studies supported scenario

(iii) with different phylogenetic groups of bacteria in each of the populations [5, 6, 31-33]. In the latter studies, the HNA population was comprised of members of the *Gammaproteobacteria*, *Bacteroidetes* and *Alphaproteobacteria* - in particular of the *Roseobacter* clade, whereas the LNA population was often dominated by the alphaproteobacterial SAR11 clade [19] or the gammaproteobacterial clade SAR86 [33]. Whether the scenarios change with changing oceanic waters is still an open question. Therefore, we analysed the phylogenetic composition of flow cytometrically sorted HNA and LNA cells from six stations in the North Atlantic Ocean with contrasting oceanographic properties between Iceland and the Azores by fluorescence *in situ* hybridization (FISH). Based on our earlier work, we hypothesized that (1) regardless of the sample origin there is little overlap in the phylogenetic composition between the LNA and HNA population, that (2) the phylogenetic composition is more diverse in the HNA than in the LNA population, and (3) that the phylogenetic composition of the LNA and HNA populations differs between the oceanic provinces investigated.

Materials and Methods

Sampling. Samples were taken onboard the research ship *Maria S. Merian* during the VISION (diVersIty, Structure, functION) cruise MSM03/1 from Reykjavik (Iceland) to the Azores (Portugal) along a transect at the 30°W longitude from 66°N to 34°N between 20th - 29th September 2006 (Fig. 1 and [11]). Seawater samples were collected from a depth of either 20 or 50 m with a sampling rosette of 20 L Niskin bottles mounted on a conductivity-temperature-depth (CTD) profiler. Replicated 1.6 mL seawater subsamples were fixed with particle-free formaldehyde solution (37% w/v, Fluka, Taufkirchen, Germany; final concentration, 1% v/v) at 2°C for 12 h and

stored frozen at −80°C. Samples for fluorescence *in situ* analyses of total bacterioplankton were fixed with particle-free formaldehyde solution (37% w/v, Fluka, Taufkirchen, Germany; final concentration, 1% v/v) for 2 hours at room temperature. Fixed samples were filtered at low vacuum onto polycarbonate filters (type GTTP; 0.2 µm pore size; 47 mm diameter; Millipore Eschborn, Germany) and afterwards washed with MilliQ to remove the remaining formaldehyde. Typically 15 – 30 ml of surface water was filtered. The filters were stored frozen at −20°C for further analyses. Total prokaryotic picoplankton and nanophytoplankton counts were taken from [11].

Sorting. The formaldehyde-fixed cells were stained with SYBR Green (conc. 1 in 1000) for 30 min prior to flow sorting with a MoFlo flow cytometer (Beckman Coulter). For excitation, an Argon ion laser (Innova-A300) was tuned to 488 nm with an output power of 500 mW. Sideward scatter (SSC) was analyzed through a 488 ± 10 nm bandpass filter, green fluorescence (FL1) of SYBR Green-stained cells was measured through a 530 ± 20 nm bandpass filter. Online analysis was done on a bivariate dot plot diagram using the Summit software V3.1 (DakoCytomation). The dotplot diagrams were used for defining sorting gates (Fig. 2). Particle-free (<0.1 µm) and autoclaved 0.1% NaCl (w/v) solution was used as a sheath fluid for sorting. The sort mode 'single one drop' was selected to get the highest sorting purity. The performance was evaluated by sorting a known number of beads onto microscopic slides which were subsequently enumerated under an epifluorescence microscope.

Catalyzed reporter deposition (CARD)-FISH. Approximately 1×10^4 cells were sorted and subsequently filtered onto 0.2 µm pore-size polycarbonate filters. CARD-FISH analyses of sorted cells and unsorted samples were done as described by Schattenhofer et al. [22] with the probe set described in Table SI 1. Cells were manually enumerated under an Axioplan II microscope (Carl Zeiss, Jena, Germany)

equipped with an HBO 100-W Hg vapour lamp, appropriate filter sets for FITC and DAPI fluorescence [10] and a 100x Plan Apochromat objective. On average >500 cells were counted per sample and probe-positive cells were presented as percentages of DAPI stained cells. The quantification of specific bacterioplankton groups in unsorted samples was done with CARD-FISH (Table SI 2) and automated cell counting as described previously [22]. Correlation analyses were done between bacterial abundances and physicochemical parameters (temperature, oxygen, NH_4^+, PO_4^{3-}, and NO_3^- + NO_2^-), chlorophyll *a* fluorescence and abundance of picoeukaryotes. Chlorophyll *a* fluorescence and nutrient data were taken from [11]. All univariated statistic analyses were performed with the software SigmaStat 3.5 (SYSTAT, California, USA).

Genome sizes. Genomic information of 227 prokaryotes, classified as marine according to the EnvO-Lite ontology, was obtained from www.megx.net [13]. The 154 genomes used in this study are the ones which have been completely sequenced, and therefore their exact genome sizes were known (Table SI 4).

Results

Environmental conditions at the sampling sites in the North Atlantic Ocean. Four major oceanographic provinces could be distinguished along the VISION cruise transect (see also [11]; Fig. 1): the Boreal Polar (BPLR), the Arctic (ARCT), the North Atlantic Drift (NADR) and the North Atlantic Subtropical East (NAST) province. Samples for the present study were derived from three of the provinces (BPLR, ARCT, NAST) and showed distinct differences in their physico-chemical and microbiological properties. Located at the boundary of BPLR and ARCT station 3 (S3) was characterized by surface waster temperature below 1 °C, low salinity (<34

psu) but relatively high concentrations of chlorophyll *a* (1.0 µg l^{-1}) and oxygen (>290 µmol). Station 6 (S6) within the ARCT province had temperature, salinity and chlorophyll *a* values of 11 °C, 35 psu and 1.7 µg l^{-1}, respectively, while oxygen concentration decreased slightly in comparison to S3 (<270 µmol). In the 4 southernmost stations of the transect in the NAST province (S16 – S19) temperature ranged between 22 – 24 °C and salinity was higher than 36 psu. In addition, lowest values in oxygen (<220 µmol) and chlorophyll *a* (0.2 – 0.3 µg l^{-1}) were detected in the NAST. The concentrations of phosphate and nitrate plus nitrite in surface waters were higher in the BPLR and ARCT with 0.4 µM and 2.9 – 5.7 µM respectively. These concentrations decreased in the NAST to values <0.6 µM while concentrations of ammonium were relatively stable between 0.2 – 0.5 µM across all stations. The total prokaryotic picoplankton was highest in surface waters of the northern stations with a maximum of 1.1 x 10^6 ml^{-1} at S4 and declined gradually southwards to 0.4 – 0.5 x 10^6 ml^{-1} at S13 to S19 (Table SI 2).

Latitudinal distribution of picoplankton groups in surface waters along the transect. The strong chemical and physical gradients along the north-south transect were also reflected in the distribution patterns of the 7 different picoplankton groups tested in surface waters. While *Bacteria* dominated the prokaryotic picoplankton (67 ± 10%, Table SI 2) at all stations (n = 19), the abundance of marine group I *Crenarchaeota* was expectedly low with 2 ± 2%. Within the domain of *Bacteria* members of the SAR11-clade comprised the most abundant fraction (27 ± 5%) followed by members of the *Bacteroidetes* (12 ± 5%). The distribution pattern of both clades was similar with highest values at northern stations, for SAR11 particularly at latitudes over 60°N (Table SI 2). *Synechococcus* were abundant in photic surface waters north of 50°N with up to 5% relative abundance, but was virtually absent further south. In contrast *Prochlorococcus* cells were detectable at S10 as the northern

most station and increased southwards up to 14% (Table SI 2). The uncultured clade SAR202 and the marine *Actinobacteria* comprised only a minor fraction of the picoplankton community in surface waters and were near the detection limit with approximately 1 ± 1% (Table SI 2). Thus these two clades were excluded from the phylogenetic analysis of flow cytometrically sorted groups of picoplankton.

Phylogenetic affiliation of flow cytometrically sorted groups of prokaryotic picoplankton. The phylogenetic composition of HNA and LNA populations was determined in water samples from the surface mixed layer from six stations (S3, S6, S16 to S19) (Fig. 3). Each population represented about half of the entire bacterioplankton community at most stations (LNA: 50.4 ± 4.4%; HNA: 49.6 ± 4.4% (n = 19), respectively, at 10 m water depth). Conversely, the sum of cell numbers calculated from the HNA and LNA fraction corroborated the total cell numbers quite well (n = 24, r^2 = 0.79). Both populations were sorted by flow cytometry and characterized by CARD-FISH to determine their phylogenetic composition (Fig. 3). Almost all sorted cells hybridised with the probe mix for *Bacteria* (HNA: 90 ± 3%, LNA: 78 ± 7%) whereas marine *Euryarchaetoa* or *Crenarchaeota* could be detected in neither of the cytometric populations.

The phylogenetic composition within the HNA population varied between the three provinces tested (Fig. 3). *Bacteroidetes* dominated the sorted HNA population (with 26% and 32%) in the BPLR and ARCT, followed by *Alphaproteobacteria (*with 19% and 25%) in which approximately half of those were represented by members of the *Roseobacter* cluster at S6 (10%) (Fig. 3). Likewise the *Gammaproteobacteria* (12% and 8%) were abundant in the BPLR and ARCT. In these two provinces only relative abundances of below 5% were detected for the uncultured *Firmicutes* clade SAR406 and the cyanobacterial genera *Synechococcus* and *Prochlorococcus* (Fig. 3). A quite different community composition of the HNA population was found in the

NAST (S16 to S19). Here the abundance of *Bacteroidetes* decreased to 3 – 10% but the majority of cells were comprised of *Prochlorococcus* cells (34 – 44%) (Fig. 3). In contrast to the rather stable abundance of *Gammaproteobacteria* (10 – 15%) the contribution of *Alphaproteobacteria* varied between 2 – 26% with a consistent fraction of *Roseobacter* (3 – 5%). Within the NAST the abundance of SAR406 decreased from 9% (S16) to 1% (S19). *Synechococcus* comprised approximately 2% at S16 to S18 (Fig. 3) as determined with probe 405Syn. The high abundance of *Prochlorococcus* within the HNA population of the NAST could be partly further characterised with subcluster-specific probes (Table SI 1). The high-light adapted *Prochlorococcus* clade I (HLI) was highly abundant between S16 to S18 (17 – 26%) but dropped to 4% at S19 (Fig. 3). The opposite distribution was obtained for the high-light adapted *Prochlorococcus* clade II (HLII). HLII showed low counts (<4%) at S16 to S18 but increased to 14% at S19. The abundance of the low-light adapted *Prochlorococcus* clade (LL) was always <2% (Fig. 3).

In contrast to the HNA population the phylogenetic composition within the LNA population was rather uniform. *Alphaproteobacteria*, and in particular the clade SAR11, dominated at all stations (Fig. 3). The LNA population of BPLR (S3) and ARCT (S6) contained besides SAR11 (with 72% and 62%) also small percentages of *Gammaproteobacteria* (<1%) and *Betaproteobacteria* (up to 4% at S6) including the betaproteobacterial clade OM43 (Fig. 3). Within the NAST the abundance of SAR11 ranged from 45% (S17) up to 74% (S19). *Betaproteobacteria* decreased to approximately 2% while *Gammaproteobacteria* increased slightly up to 3% (S18) (Fig. 3).

Discussion

The CARD-FISH analyses of unsorted surface water samples of the VISION cruise showed that the picoplankton communities were distinctly different between the different provinces examined, e.g., high *Bacteroidetes* counts in the north and low counts in the south (see also [11]). The samples were therefore sufficiently diverse to test hypotheses on a phenotypic versus genotypic differentiation of HNA and LNA cells. Even within one province the community changed like for example indicated by the change in *Prochlorococcus* ecotypes in the NAST province.

It is a main result of this study, that in all samples there was remarkably little overlap in the phylogenetic composition between the LNA and HNA. The detection of SAR11 exclusively in the LNA population (Fig. 3) confirmed previous reports from the subtropical and tropical gyres [19], and the recent report on a dominant role of SAR11 in the northern provinces [30]. Resolving the LNA population further we identified *Betaproteobacteria* as the second bacterial group present in all LNA fractions, yet in none of the HNA populations (Fig. 3). Part of the *Betaproteobacteria* were members of the uncultured OM43 group which are related to Type I methylotrophs of the *Methylophilaceae* [20] and occur commonly more in productive coastal ecosystems then in oligotrophic ocean gyres [8]. This fits well with our observed positive correlations of *Betaproteobacteria* with ammonium concentration ($r^2 = 0.91$; $p < 0.001$; Table SI 3) and nitrite plus nitrate concentrations ($r^2 = 0.90$; $p < 0.001$; Table SI 3). Surprisingly, no other phylogenetic clade could be detected in LNA populations with our set of 19 specific probes. In previous studies a high abundance (>10%) of the gammaproteobacterial clade SAR86 was reported in the LNA fraction of prokaryotic picoplankton in coastal seas [33, 34]. However, in this study the SAR86 clade was found exclusively in the HNA population sorted from the

two productive provinces BPLR and ARCT, thus confirming their preference for highly productive environments like upwelling regions [26] or the coastal North Sea [4]. One possible reason for this discrepancy could be slight variations in the manual definition of sorting gates on different flow cytometers with different fluorescent DNA stains between different studies (see [33]). Another reason might be the growth stages of SAR86 cells depending on the environment they are living in [1]. Fuchs and co-workers retrieved SAR86 sequences from sorted cells of the Arabian Sea from both the LNA and HNA fraction, but their respective phylogenetic affiliation was different [5]. It can be speculated that some SAR86 sub-clades might have a small genome and are consequently detected in the LNA fraction, while others having a larger genome are falling into the HNA settings. Alternatively, the genome copy number of SAR86 might vary depending on growth conditions. This needs to be clarified in another study focussing on the SAR86 clade.

Our assessment of the HNA population resulted in signals for 13 out of 19 probes used for targeting clades at different phylogenetic levels (Table SI 1). In contrast, only 5 probes gave signals with sorted LNA cells thus confirming our hypothesis that the HNA fraction is more diverse than the LNA fraction [5, 31, 34]. Next to the more prominent groups like *Roseobacter* and *Prochlorococcus*, members of uncultured groups like the SAR324 (data not shown) and SAR406 were detected in the HNA fraction indicating that the genomes of these yet uncultured clades are rather large. Furthermore, the HNA community varied between the stations along the transect in the North Atlantic Ocean. The high abundance of *Bacteroidetes* (genome size ranges from 3 – 6 Mbp, Table SI 4) in HNA populations from the productive BPLR and ARCT provinces (Fig. 3) is likely due to their role as consumers of algae-derived polymeric substances in this area of the Northern Atlantic Ocean [11]. Positive correlations of *Bacteroidetes* in the HNA group with e.g. chlorophyll *a* (r^2 =

0.94; p = 0.001), phosphate (r^2 = 0.98; p <0.001) or with picoeukaryotic phytoplankton (r^2 = 0.92; p <0.003) again corroborate a preference of *Bacteroidetes* for nutrient-rich water masses (Table SI 3). The *Roseobacter* also showed a higher abundance in sorted HNA populations at one of the stations in the northern provinces (Fig. 3) coinciding with the highest values of chlorophyll *a* found along the transect (>1.5 µg l^{-1}). The marine *Roseobacter* clade comprises phylogenetically diverse and physiologically versatile bacterial species [3] and is commonly found abundant in costal areas or shelf regions associated with phytoplankton blooms [29, 32]. The genome sizes of cultured *Roseobacter* representatives range from 3.1 to 5.5 Mbp.

Flow cytometric analyses showed an almost perfect 1:1 – split of the picoplankton community into LNA and HNA populations along the entire transect. However, from there it became evident, that the CARD-FISH counts of flow cytometrically sorted populations deviate from the quantifications of the entire community. For example Cren- and Euryarchaeota could not be detected in any of the sorted populations, although they were present in up to 2% relative abundance in the unsorted sample. One reason could be a detection limit for CARD-FISH on sorted populations which lies at the level of ~1% relative abundance. Another reason could be cell loss during flow cytometric sorting. Together with subjective gating of flow cytometric populations this might account for the observed deviations of abundance of sorted and unsorted picoplankton clades.

Genome size estimation of LNA and HNA populations. A simple interpretation of the conspicuous bimodal distribution pattern in DNA-scatter dotplot diagrams would be that the two flow cytometric populations consist of cells in different stages of DNA replication (n versus 2n) due to cellular division. We can exclude this explanation for our samples because the cytometric populations contain different microbial clades with almost no overlap and the average DNA content of

HNA cells is 3.5 times that of LNA. Consequently different genome sizes are the most likely reason for the bimodal distribution pattern (Fig. 2). Islas and coworkers [12] reported previously a bimodal distribution of the genome sizes of 641 free living prokaryotes. In contrast, our own analyses of the genome sizes of 153 cultivated *Bacteria* and 1 *Archaea*, manually classified as marine according to the EnvO-Lite ontology [13], show little evidence for bimodality (Fig. SI 1; Table SI 4). In our samples the bimodal distribution in DNA-scatter dotplots might rather be a function of genome sizes and of abundant clades (Fig. 2). It is well known that members of the dominant SAR11 clade have small genome sizes ranging between 1.3-1.5 Mbp [9]. Another clade with a cultured representative having a genome size around 1.3 Mbp is the OM43 clade [8]. We found OM43 in small amounts only in LNA samples from the BPLR and ARCT supporting our hypothesis. We have not yet identified the remaining 10 – 25% of *Bacteria*, which have similar small genomes as SAR11 and OM43. The slightly higher hybridisation rates with ALF968 over SAR11-441 suggest that there might be another - yet unknown - small genome-sized alphaproteobacterial group in the LNA population, which were not picked up by the probes used for CARD-FISH. In deeper water layers candidate organisms with small genome sizes are members of the marine group I *Crenarchaeota*. *Nitrosopumilus maritimus*, the only cultured marine group I *Crenarchaeota*, has a genome size of 1.6 Mbp [28]. We failed to find them in sorted fractions from surface waters (see above), but we could detect them in higher amounts in the LNA from test samples originating from deeper water layers (data not shown).

Along the transect the ratio between the mean DNA fluorescence of the HNA population and the mean DNA fluorescence of the LNA population was approximately 3.5 ± 0.2 (n = 6). This value multiplied with the genome size of 1.3 Mbp of the dominant cell type SAR11 results in a value of 4.6 Mbp, which lies indeed

in the range of a "standard genome size" of 3 – 6 Mbp of many of the phyla detected in the HNA population (Table SI 4). However, this hypothesis is challenged by the flow cytometric detection of *Prochlorococcus* spp. in the HNA fraction [35]. With a genome size range of 1.6 – 2.4 Mbp [21] *Prochlorococcus* spp. should theoretically be located between the LNA and HNA populations based on their DNA fluorescence. However for most of the samples analysed by flow cytometry the DNA-SYBR Green fluorescence rather suggested a genome size of around 4 Mbp for *Prochlorococcus*. One explanation for the detection of *Prochlorococcus* in the HNA population might be that the autofluorescence of photosynthetic pigments adds to the SybrGreen fluorescence of DNA. Although the main fluorescence emission is in the red, the pigments contribute considerably to the fluorescence of *Prochlorococcus* cells in the green channel and thus adding to the DNA conferred fluorescence and thereby potentially shifting the cells into the HNA population (see also [16]). Other phototrophic microorganisms with genome sizes below 3 Mbp, e.g. *Synechococcus*, might also be affected by their autofluorescence and thus shifted into the HNA population, although we could not detect such an effect in our dataset. We can also not exclude that genome copy numbers are higher in *Prochlorococcus* (e.g. [27])

Conclusion

Of the scenarios summarised by Bouvier and coworkers [2] our results suggest that scenario (iii) – LNA and HNA populations are composed of distinct phylogenetic clades – is the dominant one in large areas of the open ocean even across strong physico-chemical gradients. In rare cases like blooming situations scenario (i) – cells start growing and develop from LNA to HNA cells – might prevail for e.g. SAR86 when they change their phenotype and multiply their genomes. None of the scenarios fit adequately for *Prochlorococcus* and other pigmented microorganisms.

From an ecological perspective the stable distinction into two different DNA-containing populations might reflect fundamentally different life strategies. Members of the HNA population with large genomes are theoretically able to cope with a wide variety of environmental conditions, while LNA populations with small streamlined genomes have a rather limited genetic repertoire and occupy narrow ecological niches. Hence, the LNA and HNA concept may remain useful in interpreting the ecological role of each population.

Acknowledgement

We thank the Captain, Officers and Crew of the FS *Maria S. Merian* (cruise MSM03/01) for their help during the cruise. We are grateful especially to Rudolf Amann and three anonymous reviewers for their helpful comments on earlier versions of this manuscript. This study was funded by the Max Planck Society.

References

[1] Bouvier, T., del Giorgio, P.A. and Gasol, J.M. (2007). A comparative study of the cytometric characteristics of high and low nucleic-acid bacterioplankton cells from different aquatic ecosystems. Environ. Microbiol. 9, 2050-2066.

[2] Brussaard, C.P.D., Marie, D. and Bratbak, G. (2000). Flow cytometric detection of viruses. J. Virol. Methods 85, 175 - 182.

[3] Buchan, A., Gonzalez, J.M. and Moran, M.A. (2005). Overview of the Marine Roseobacter Lineage. Appl. Environ. Microbiol. 71, 5665-5677.

[4] Eilers, H., Pernthaler, J., Peplies, J., Gloeckner, F.O., Gerdts, G. and Amann, R. (2001). Isolation of novel pelagic bacteria from the German Bight and their seasonal contributions to surface picoplankton. Appl. Environ. Microbiol. 67, 5134-5142.

[5] Fuchs, B.M., Woebken, D., Zubkov, M.V., Burkill, P.H. and Amann, R. (2005). Molecular identification of picoplankton populations in contrasting waters of the Arabian Sea. Aquat. Microb. Ecol. 39, 145-157.

[6] Fuchs, B.M., Zubkov, M.V., Sahm, K., Burkill, P.H. and Amann, R. (2000). Changes in community composition during dilution cultures of marine bacterioplankton as assessed by flow cytometric and molecular biological techniques. Environ. Microbiol. 2, 191-202.

[7] Gasol, J.M., Zweifel, U.L., Peters, F., Fuhrman, J.A. and Hagstrom, A. (1999). Significance of size and nucleic acid content heterogeneity as measured by flow cytometry in natural planktonic bacteria. Appl. Environ. Microbiol. 65, 4475-4483.

[8] Giovannoni, S.J. et al. (2008). The small genome of an abundant coastal ocean methylotroph. Environ. Microbiol. 10, 1771-1782.

[9] Giovannoni, S.J. et al. (2005). Genome streamlining in a cosmopolitan oceanic bacterium. Science 309, 1242-1245.

[10] Glöckner, F.O., Fuchs, B.M. and Amann, R. (1999). Bacterioplankton composition in lakes and oceans: a first comparison based on fluorescence *in situ* hybridization. Appl. Environ. Microbiol. 65, 3721-3726.

[11] Gomez-Pereira, P.R., Alonso, C., Oliver, M., van Beusekom, J. and Fuchs, B.M. (2010). Distinct flavobacterial communities in contrasting water masses of the North Atlantic Ocean. ISME Journal

[12] Islas, S., Becerra, A., Luisi, P.L. and Lazcano, A. (2004). Comparative genomics and the gene complement of a minimal cell. Origins of Life and Evolution of Biospheres 34, 243-256.

[13] Kottmann, R., Kostadinov, I., Duhaime, M.B., Buttigieg, P.L., Yilmaz, P., Hankeln, W., Waldmann, J. and Glöckner, F.O. (2010). Megx.net: integrated database resource for marine ecological genomics. Nucleic Acids Res. 38, D391-395.

[14] Lebaron, P., Servais, P., Agogue, H., Courties, C. and Joux, F. (2001). Does the high nucleic acid content of individual bacterial cells allow us to discriminate between active cells and inactive cells in aquatic systems? Appl. Environ. Microbiol. 67, 1775-1782.

[15] Lebaron, P., Servais, P., Baudoux, A.-C., Bourrain, M., Courties, C. and Parthuisot, N. (2002). Variations of bacterial-specific activity with cell size

and nucleic acid content assessed by flow cytometry. Aquat. Microb. Ecol. 28, 131-140.

[16] Li, W.K.W., Jellett, J.F. and Dickie, P.M. (1995). DNA distributions in planktonic bacteria stained with TOTO or TO-PRO. Limnol. Oceanogr. 40, 1485-1495.

[17] Longhurst, A. (1998) Ecological Geography of the Sea, Academic Press. New York.

[18] Marie, D., Partensky, F., Jacquet, S. and Vaulot, D. (1997). Enumeration and cell cycle analysis of natural populations of marine picoplankton by flow cytometry using the nucleic acid stain SYBR Green I. Appl. Environ. Microbiol. 63, 186-193.

[19] Mary, I., Heywood, J.L., Fuchs, B.M., Amann, R., Tarran, G.A., Burkill, P.H. and Zubkov, M.V. (2006). SAR11 dominance among metabolically active low nucleic acid bacterioplankton in surface waters along an Atlantic meridional transect. Aquat. Microb. Ecol. 45, 107-113.

[20] Morris, R.M., Longnecker, K. and Giovannoni, S.J. (2006). *Pirellula* and OM43 are among the dominant lineages identified in an Oregon coast diatom bloom. Environ. Microbiol. 8, 1361-1370.

[21] Rocap, G. et al. (2003). Genome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature 424, 1042-1047.

[22] Schattenhofer, M., Fuchs, B.M., Amann, R., Zubkov, M.V., Tarran, G.A. and Pernthaler, J. (2009). Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. Environ. Microbiol. 11, 2078 - 2093.

[23] Servais, P., Casamayor, E.O., Courties, C., Catala, P., Parthuisot, N. and Lebaron, P. (2003). Activity and diversity of bacterial cells with high and low nucleic acid content. Aquat. Microb. Ecol. 33, 41-51.

[24] Servais, P., Courties, C., Lebaron, P. and Troussellier, M. (1999). Coupling bacterial activity measurements with cell sorting by flow cytometry. Microb. Ecol. 38, 180-189.

[25] Sherr, E.B., Sherr, B.F. and Longnecker, K. (2006). Distribution of bacterial abundance and cell-specific nucleic acid content in the Northeast Pacific Ocean. Deep Sea Res. Part I 53, 713-725.

[26] Suzuki, M.T., Preston, C.M., Chavez, F.P. and DeLong, E.F. (2001). Quantitative mapping of bacterioplankton populations in seawater: field tests across an upwelling plume in Monterey Bay. Aquat. Microb. Ecol. 24, 117-127.

[27] Vaulot, D., Marie, D., Olson, R.J. and Chisholm, S.W. (1995). Growth of *Prochlorococcus*, a Photosynthetic Prokaryote, in the Equatorial Pacific Ocean. Science 268, 1480-1482.

[28] Walker, C. B.; de la Torre, J. R.; Klotz, M. G.; Urakawa, H.; Pinel, N.; Arp, D. J.; Brochier-Armanet, C.; Chain, P. S. G.; Chan, P. P.; Gollabgir, A.; Hemp, J.; Hügler, M.; Karr, E. A.; Könneke, M.; Shin, M.; Lawton, T. J.; Lowe, T.; Martens-Habbena, W.; Sayavedra-Soto, L. A.; Lang, D.; Sievert, S. M.; Rosenzweig, A. C.; Manning, G. and Stahl, D. A. *Nitrosopumilus maritimus* genome reveals unique mechanisms for nitrification and autotrophy in globally distributed marine crenarchaea Proc. Natl. Acad. Sci. U. S. A., 2010, 107, 8818 -8823.

[29] West, N.J., Obernosterer, I., Zemb, O. and Lebaron, P. (2008). Major differences of bacterial diversity and activity inside and outside of a natural iron-fertilized phytoplankton bloom in the Southern Ocean. Environ. Microbiol. 10, 738-756.

[30] Wietz, M., Gram, L., Jørgense, B. and Schramm, A. (2010). Latitudinal patterns in the abundance of major marine bacterioplankton groups. Aquat. Microb. Ecol. 61, 179-189.

[31] Zubkov, M.V., Allen, J.I. and Fuchs, B.M. (2004). Coexistence of dominant groups in marine bacterioplankton community - a combination of experimental and modelling approaches. J. Mar. Biol. Assoc. U. K. 84, 519-529.

[32] Zubkov, M.V., Fuchs, B.M., Archer, S.D., Kiene, R.P., Amann, R. and Burkill, P.H. (2001). Linking the composition of bacterioplankton to rapid turnover of dissolved dimethylsulphoniopropoinate in an algal bloom in the North Sea. Environ. Microbiol. 3, 304-311.

[33] Zubkov, M.V., Fuchs, B.M., Burkill, P.H. and Amann, R. (2001). Comparison of cellular and biomass specific activities of dominant bacterioplankton groups in stratified waters of the Celtic Sea. Appl. Environ. Microbiol. 67, 5210-5218.

[34] Zubkov, M.V., Fuchs, B.M., Tarran, G.A., Burkill, P.H. and Amann, R. (2002). Mesoscale distribution of dominant bacterioplankton groups in the northern North Sea in early summer. Aquat. Microb. Ecol. 29, 135-144.

[35] Zubkov, M.V., Sleigh, M.A., Tarran, G.A., Burkill, P.H. and Leakey, R.J.G. (1998). Picoplanktonic community structure on an Atlantic transect from 50°N to 50°S. Deep Sea Res. Part I 45, 1339-1355.

Figures

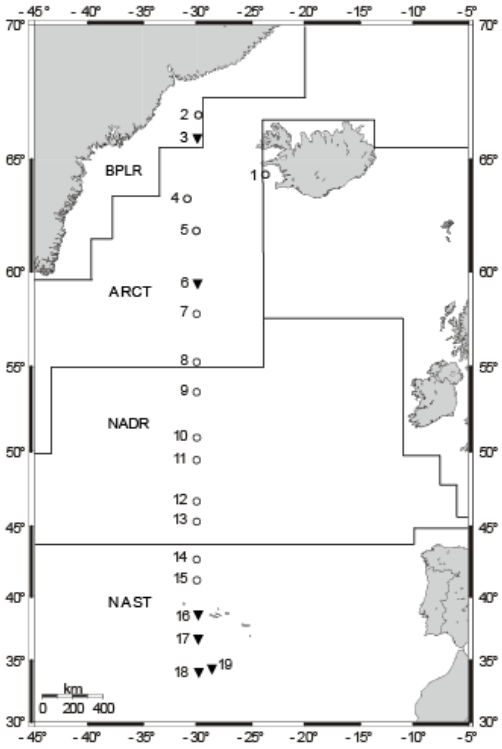
Fig 1: VISION cruise track from Reykjavik, Iceland to the Azores, Portugal during September 2006. Boundaries and abbreviations for the individual oceanic provinces *sensu* Longhurst [17]: Boreal Polar (BPLR), Arctic (ARCT), North Atlantic Drift (NADR) and North Atlantic Subtropical East (NAST) province. Triangles indicate stations for flow cytometric sorting. For more details see also [11].

Fig 2: Example of a dot plot diagram of the flow cytometric analysis of a picoplankton sample from surface waters at station 18. Gates for flow cytometric sorting of HNA and LNA populations are depicted by black boxes.

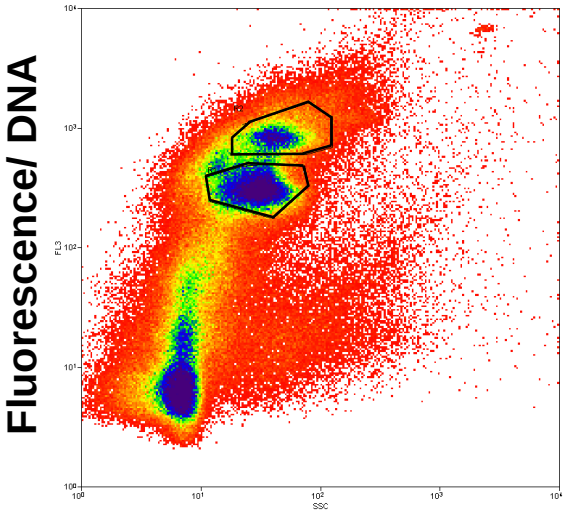
Fig 3: Relative abundance (% DAPI counts) of prokaryotic picoplankton groups within unsorted and sorted (HNA and LNA) samples of different oceanic provinces in the North Atlantic Ocean.

Figure 1-3 , Suppl Fig 1

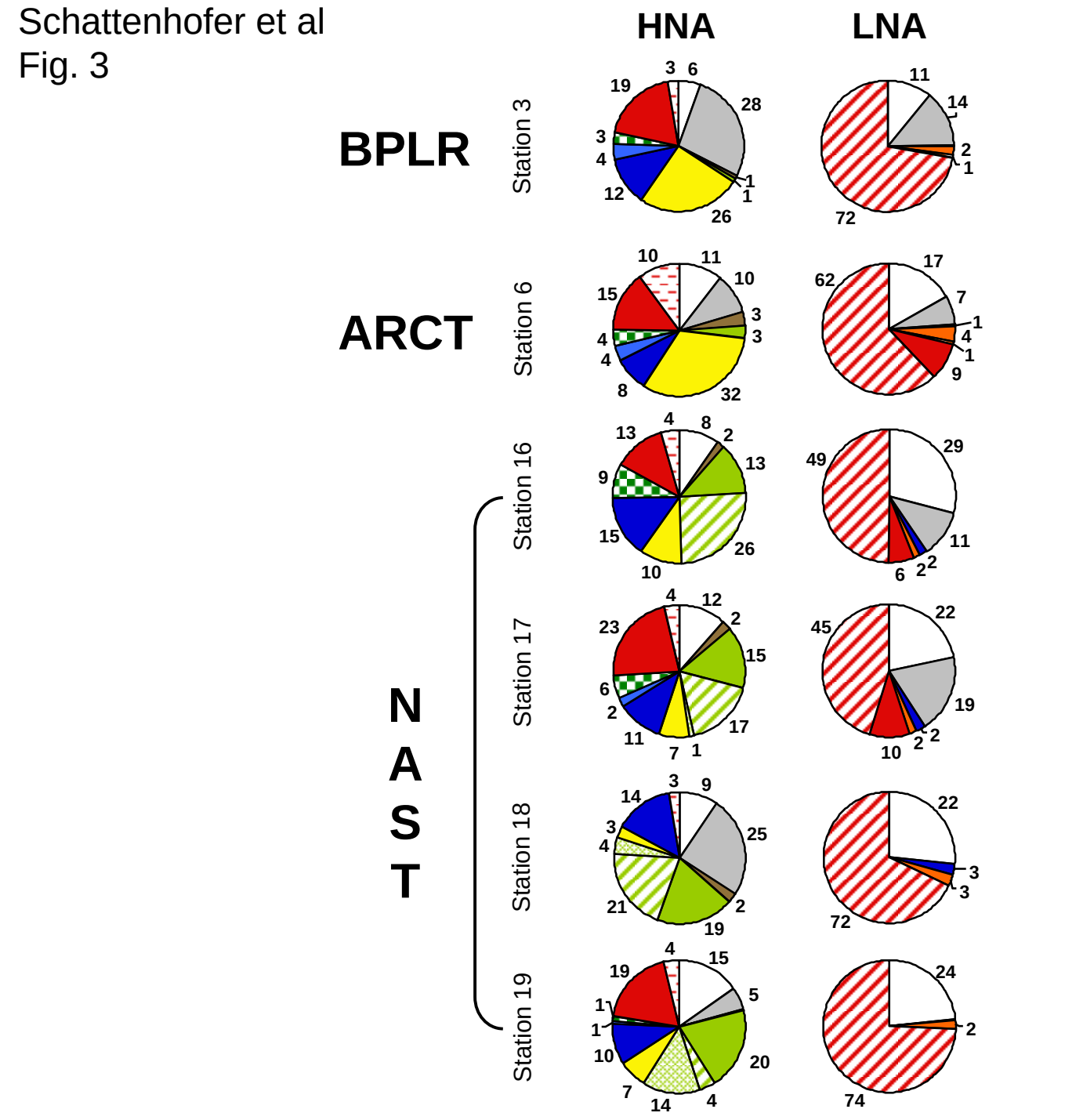
Schattenhofer et al
Fig. 1



Schattenhofer et al
Fig. 2



Schattenhofer et al
Fig. 3



CHAPTER 3

SUMMARY OF PUBLICATIONS

The aim of this thesis was to bring the field of marine ecological genomics forward by (1) improving integration of ecological and molecular data through the development of the megx.net platform and (2) using the platform for ecogenomic analysis. The resulting scientific publications can be divided into two sections. The first one deals with the megx.net platform as an enabling technology for marine ecological genomics. The second one presents three studies which make use of the resources offered by megx.net.

3.1 Enabling Technology for Marine Ecological Genomics: Megx.net

Considering the relatively small number of marine (meta)genomic samples at the time the megx.net portal was initiated [Lombardot et al., 2006], the manual effort needed to gather and process the data was manageable. With interest in marine microbes rising and sequence data accumulating exponentially, this is no longer the case. The major data contributions by the GOS Expedition [Venter et al., 2004, Rusch et al., 2007] and the Marine Microbes Initiative [7] set a landmark in the field of marine genomics. But this is by far not the peak of genomic data coming in. In light of these rapid changes, the megx.net portal was updated to meet the evolving demand for advanced integration of environmental and genomic data from the marine realm [Kottmann et al., 2010]. Some of the key updates include:

- A new relational data model
- A data update to all currently sequenced genomes (complete and

[7] http://www.moore.org/marine-micro.aspx

drafts), the GOS metagenome and sequenced marine viruses

- Integration of environmental parameters from the World Ocean Atlas, World Ocean Database [Boyer et al., 2006] and SeaWIFS chlorophyll α
- A manual classification of genomic samples according to EnvO-Lite [Hirschman et al., 2008]
- MIGS/MIMS/MIMARKS compliance [Field et al., 2008,Yilmaz et al., 2011b]
- A scalable, high-throughput implementation of Geographic-BLAST
- Web Service access

The main goal of the megx.net project is to gather data from publicly available resources, integrate it as best as possible and bring it back to the public domain, together with the appropriate tools to access and analyze it. The central concept of integration is that every location on Earth can be uniquely identified by its longitude and latitude (*x, y*) and a sample from that location can be uniquely identified by adding the depth and time of sampling (*z, t*). Samples are referred to as georeferenced if at least the longitude and latitude are known.
The megx.net portal can be generally divided into two parts. The relational database back-end (MegDb) serves for storing and integrating the data. The web-based front-end provides access to it and the appropriate analysis tools. MegDb was designed to hold data of every major aspect of ecological genomics - from sample description and on-site measurements of environmental parameters through sequencing procedures to bioinformatic analysis. The new data model allows MegDb to scale in size and to accommodate further aspects of sequence-centered marine ecology on demand.
Besides the comprehensive sequence data update, megx.net profits from integration of high-quality environmental data. Nine environmental parameters are available for every 1 degree grid of the ocean in 33 standard depths and over several different time spans. For other depths, inverse distance interpolation is employed. A numeric measure for environment stability was introduced [Kostadinov et al.,]. The

added value of interpolated data for marine ecological genomics is discussed in section 2.3. Environmental data can be visualized directly on the Genes Mapserver (Figure 3.1). Megx.net was the first resource to offer classification of microbial genomes according to the EnvO-Lite ontology [Hirschman et al., 2008]. It allows to select all genomes coming from a marine environment in a single mouse-click. The megx.net project also pioneered data compliance to the MIGS/MIMS/MIMARKS [Field, 2008, Yilmaz et al., 2011b] standards and the use of the GCDML exchange format [Kottmann et al., 2008].

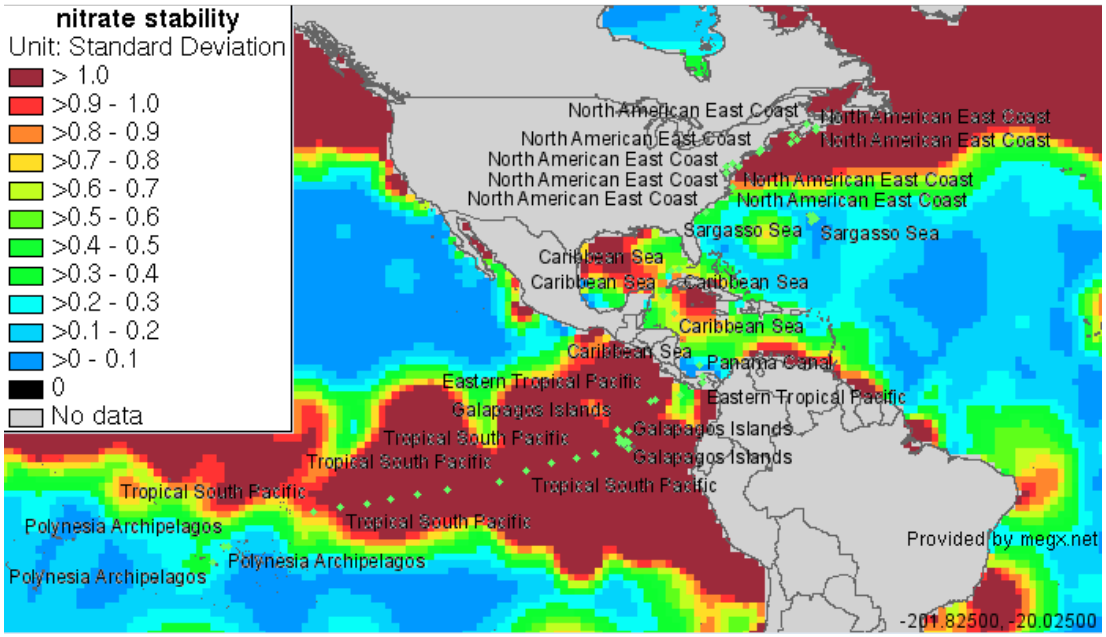


Figure 3.1: Nitrate stability at the surface visualized with the Genes Mapserver. The green dots are samples from the GOS Expedition. Image from `www.megx.net` (modified)

The tools to access and analyze the megx.net data were improved as well, all environmental and stability data can be directly visualized on the Genes Mapserver, together with different sample types or the results of a Geographic-BLAST. The new Geographic-BLAST implementation makes use of the MegDb back-end for asynchronous job submission and retrieval. This means a user can return at any time to view their results. The BLAST search is performed in a cluster environment with load balancing ensuring performance and scalability.

Programmatic access to megx.net data in the form of Web Services (WS) is now available. WS facilitate a collective view of data from different sources. Megx.net and the SILVA ribosomal database [Pruesse et al., 2007] are an example for such collaboration. Georeferenced rRNA sequences from any given sampling site in megx.net can be pre-selected on the SILVA web page. Vice-versa, the SILVA website provides a direct link to interpolated environmental data for georeferenced rRNA sequences. This simple mechanism alone allows multiple entry points of research: a phylogenetic one through SILVA and an environmental one through megx.net. With the same ease, WS can be used to directly visualize data from different resources on the same interface. An example implementation in megx.net provides descriptors of the MIGS/MIMS/MIMARKS terms [Field, 2008, Yilmaz et al., 2011b]. This WS is already in use by CDinFusion, a tool for contextual data enrichment of sequence FASTA files [Hankeln et al.,].

The research possibilities megx.net now offers have already been used to study the ecology of marine microorganisms. Some of the analyses address topics that were put forward by the original megx.net project [Lombardot et al., 2006]. For example, quantifying the environmental adaptation of microbial transcription factor repertoire (Section 3.2.1) and gaining an ecological perspective on protein domains of unknown function (Section 3.2.2). Integrated genomics data from megx.net was also used to enhance the interpretation of results from classical molecular techniques like FISH (Section 3.2.3).

In summary, megx.net was transformed into a robust and flexible platform for marine ecological genomics. With its focus on environmental data integration, it is still a unique resource for marine ecological genomics.

3.2 Metadata-Supported Ecogenomics of Marine Microbes

3.2.1 Quantifying the Effect of Environment Stability on the Transcription Factor Repertoire of Marine Microbes

This study is one of the first ecological genomic studies, fully utilizing the megx.net resources. It was triggered by reports that the complex lifestyle of microbes from dynamic coastal areas is mirrored in their gene content, especially in the control of gene expression [Cases et al., 2003, Palenik et al., 2006, Yooseph et al., 2010]. It represents the most comprehensive use of interpolated data for ecological interpretation of sequences so far. Interpolated data was shown to have statistically significant predictive power for *in situ* measurements. The work builds up on the usage of imputed data. Imputation describes the replacement missing data points by a meaningful value, often calculated from the available points. Figure 3.2 shows the long term perspective interpolated data offers in comparison to single *in situ* measurements. Testing the imputation quality of interpolated data revealed two important limitations. Firstly, no reliable interpolations can be made for undersampled regions. This issue can be resolved by increasing the integration of on-site *in situ* measurements. Secondly, a discrepancy often exists between implicit and explicit knowledge. An example for implicit knowledge is the categorical description like 'hypersaline mangrove forest'. Explicit knowledge would be numeric data supporting this description. A perfect example for the discrepancy between the two is the sample GS033.

The variation in environmental parameters can influence the TF content of microbial communities significantly. Still, about 60% of the variation in the total TF content remained unexplained. Whether the remaining variation is due to yet untested environmental factors or completely different causal agents remains to be seen. To address this question one could repeat the analysis with an even more comprehensive set of environmental parameters and further contextual data. Unexpected factors that influence transcription control could also be

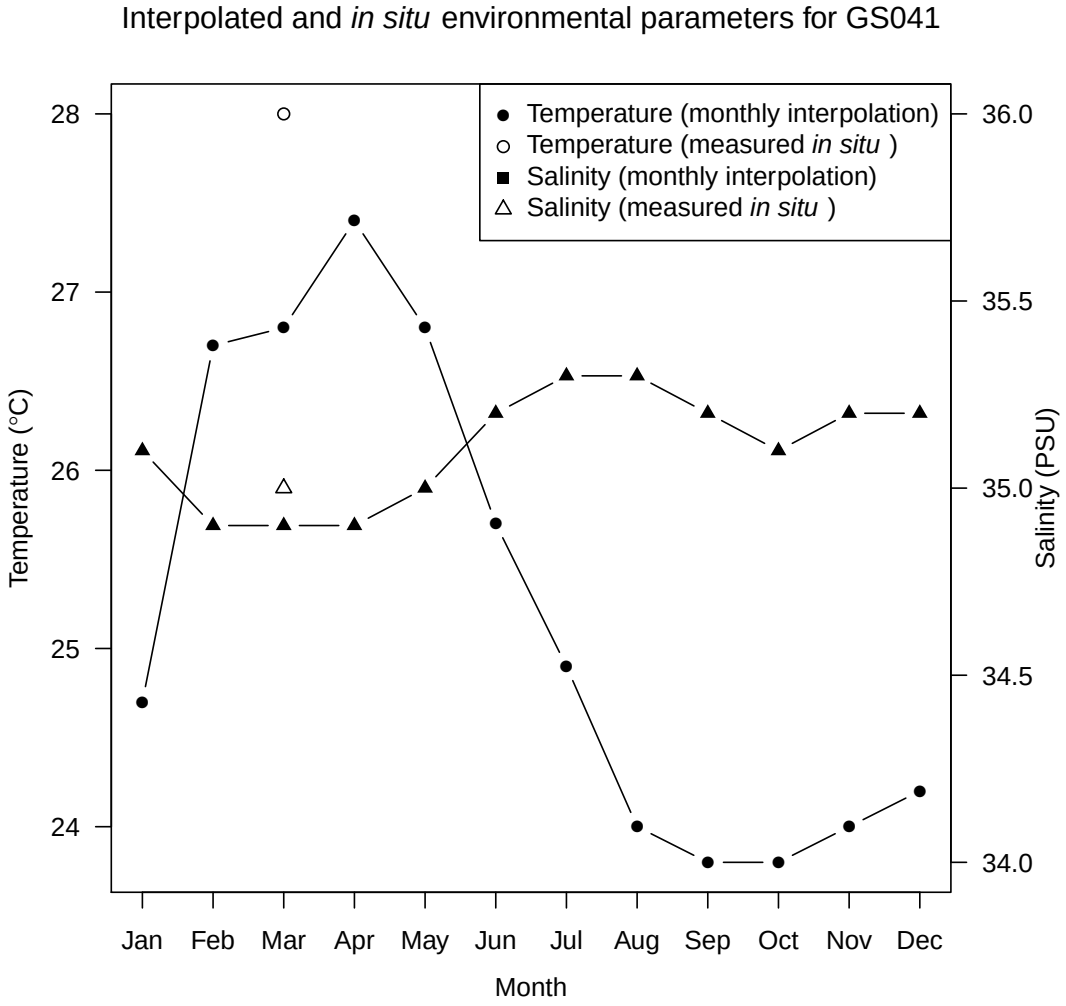


Figure 3.2: Temperature and salinity for GOS sample GS041. In situ measurements (unfilled symbols) from on board the Sorcerer II give a momentary glimpse into the environment at a given sampling site. The monthly interpolations (filled symbols) provide a basis for comparison to averages based previous measurements. The lines were added to better visualize the trends.

identified this way. Furthermore, any influence of taxonomic and phylogenetic differences on the variation in TF content should also be investigated for completeness.
The approach used here is generally applicable to other genomic features like sulfatase genes and glycosyl hydrolase genes. The resources offered by megx.net and the supplemental R code from this publication allow to easily reproduce and modify the analysis. However, the specifics of the genes, organisms, habitats, environmental parameters and statistical methods have to be taken into account. Therefore, this kind of analysis cannot be turned into a pipeline, which delivers a definitive answer at the click of a button. Consequently, this study offers a guideline for exploring the relationships between microbial gene content and the environment.

3.2.2 Ecological perspectives on domains of unknown function: a marine point of view

This study demonstrates an application of protein domain detection in high-throughput metagenomic data, integrated environmental parameters and graph theory to generate functional hypotheses for protein domains of yet unknown function. A set of co-occurring domains could be loosely described as a microbial photoreactivity module. Their abundances structure the samples along the chlorophyll concentration gradient.
The approach presented here can be adapted to investigate other sets of protein domains. Domains with known function could be tested for new co-occurrence patterns, which could form new hypotheses about unknown interactions between them. The initial exploratory phase should be followed by an in-dept analysis of a selected module. For example, the gene-neighborhood organization of the module could be investigated in completely sequenced genomes and any conserved structures could be tested for co-occurrence in metagenomic data.
The study addresses a central theme of the megx.net project: attempting to functionally characterize genes by using their environmental context. The importance of such approaches is rising. A large portion of the predicted genes in prokaryotic genomes lack functional annota-

tion based on *in silico* methods and wet-lab experiments are costly in both time and resources [Karaoz et al., 2004]. The rapidly increasing amount of sequence data magnifies this problem tremendously. Novel, high-throughput solutions like the one presented here must be sought.

3.2.3 Phylogenetic Characterisation of Picoplanktonic Populations with High and Low Nucleic Acid Content in the North Atlantic Ocean

This study focused on the differences in phylogenetic composition between marine prokaryotic picoplankton with high and with low nucleic acid content. It is an example of the use of integrated metadata in a classical study of molecular ecology. In this case the megx.net resources were used to collect the genome sizes of marine bacteria. Although the genome size is an easy to access statistic in many public resources (e.g. GOLD [Liolios et al., 2010]), identifying a marine bacterium is no trivial task. Genomic data annotated with an appropriate ontology like EnvO-Lite [Hirschman et al., 2008] provides an easy and comparable solution.

CHAPTER 4

OUTLOOK

4.1 From enabling technology to enabling design

The field of genomics is a quickly evolving one. While we have struggled to tackle one big piece of the data pie, the GOS Expedition[8] recently finished another impressive voyage. The route covered transects in the Northern Atlantic Ocean, the Mediterranean, the North Sea and the Black Sea. The GOS Expedition is not the sole producer of high-throughput genomic and contextual data from the marine environment. Large-scale projects like TARA Oceans[9], Malaspina [10] and the Earth Microbiome Project [11] generate comprehensive datasets about the marine microbial communities in their environment. This means many more new sequences lay in store for us in the near future. The new design of the megx.net platform allows it to tackle an ever greater amount and complexity of data in the future. Besides accommodating the upcoming data, the megx.net portal can offer improved analytical tools and metadata. The quality of interpolations depends massively on the number and quality of *in situ* measurements they are based on [Kostadinov et al.,]. Integrating further environmental data and including it into the interpolation procedure could offer a boost in the accuracy of interpolated environmental parameters. An appropriate candidate source for such data is for example PANGAEA[12]. Further, in order for analysis to keep pace with the rising amounts of data and its complexity, the tools to do so must evolve. The megx.net portal was taken a step in the right direction with its new implementation

[8] http://www.jcvi.org/cms/research/projects/gos/overview/
[9] http://oceans.taraexpeditions.org
[10] http://www.expedicionmalaspina.es
[11] http://www.earthmicrobiome.org/
[12] http://www.pangaea.de/

of Geographic-BLAST. A logical extension to that would be a tool to detect the possible protein domains (based on Pfam) in a sequence of interest (Geographic-HMMER). The results could be visualized on the Genes-Mapserver as distributions of the found domains across the ocean. Pre-computed frequencies of Pfam domains could be visualized per sampling site. This would revive a feature of the initial megx.net website which was in the meantime difficult to support.
The Internet plays a key role in the development of many scientific fields. For biology, the greatest impact is the faster transfer of data and easier communication between researchers. The megx.net project plans to increase its data quality through community annotation. A major downside of semi-automatic data transfer and integration is that many errors which would be obvious to a human remain unknown to a machine. A real-life example from the megx.net project is a sample which was labeled as Mediterranean but according to its coordinates was taken from the Atlantic ocean. The mistake turned out to be a missing minus sign in the longitude value. Such an error is easy to spot when looking at a label on a map, but difficult to identify automatically. Therefore, the megx.net would like to offer its users an intuitive way to increase the quality and the added-value of the data. This should be done by applying a social web feature commonly used in platforms like web-based video and radio platforms (e.g. `http://last.fm`, `http://www.youtube.com`). The users will be offered the possibility to add the so called 'tags' and 'flags' to any piece of data which is shown on the web page. In our terminology, a tag is a short description defined by the user. It would help the users identify data which is of interest to them and create their own dataset. Users will be allowed to browse the data according to tags added by all visitors, or only by themselves (requires login facility). A flag is a pre-defined tag which will be used to identify how good a piece of data is. In the scenario above, any user that recognized the mistake in the data, could mark it as wrong by setting an appropriate flag to it. This serves two purposes. Firstly, users responsible for data management and quality (data curators) can identify problems and deal with them. Secondly, all users will be aware of the quality of a certain piece of data, if it is flagged as "correct" by several people. This principle of community annotation is gaining importance, although its efficacy and

quality are often questioned. Successful implementations include the Annotathon system, which is used to annotate metagenomes as part of the coursework of undergraduate students [Hingamp et al., 2008]. The approach megx.net plans to use is minimalistic and intuitive, which should prompt user acceptance. The community annotation feature will be tested internally as an internal curation interface and in a worse case scenario be used as such in the future. The database modules that were already developed for that purpose are done in a way so they can deliver the same functionality also outside the megx.net environment. Megx.net provides not only a platform for marine ecological genomics but also a template for working with next-generation biological data.

4.2 Technology- and hypothesis-driven marine genomics

It is often argued whether the main driving force behind scientific progress is technology or the question it serves to answer. It is however a race of constant overtaking and mutual promotion. A hypothesis without the tools to test it will prompt their development. Technologies that deliver new spectra of data pave the way for completely new questions to be asked.

The first GOS Expedition was mostly technologically driven. However, the benefits of having extremely large metagenomic datasets for studying marine microbial communities quickly became clear. Using contextual data to improve the interpretation was proven to be successful [Gianoulis et al., 2009, Yooseph et al., 2010, Kostadinov et al.,]. The lesson was obviously learned in time, because the second GOS cruise collected more environmental data than the first. Additionally, large-scale projects like TARA Oceans, Malaspina and the Earth Microbiome Project try to focus their efforts on specific questions. Malaspina for example deals mainly with the biodiversity of the deep ocean and effects of climate change. These projects will produce not only large amounts of metaomics data, but probably an equal or hopefully larger amount of contextual data. The ecogenomic studies possible with megx.net were so far limited by the lack or inconsistency of contextual data. There-

fore, the input of TARA Oceans or of the Earth Microbiome Project will be very welcome. Additionally, dedicated marine observation networks like the European Marine Observation and Data Network (EMODNET) and the Ocean Observatories Initiative (OOI) collect oceanographic data on a fine scale over long periods of time. Such data could improve the interpolations megx.net offers tremendously, especially in the coastal areas. To enable innovative ecogenomic techniques, megx.net will have to either be able to host the data and integrate it internally or exchange it on the fly via Web Services. Having properly standardized data is a prerequisite for success. Therefore, the activities of the GSC will most likely have the strongest impact on megx.net and all other marine ecological resources.

Marine microbial genomics is a field that entirely depends on data. Even the most tightly focused hypothesis cannot be proven without data. And incidentally, data taken for one purpose can sometimes serve another. Therefore, an integration infrastructure like megx.net offers the perfect opportunity to develop new hypothesis but also to re-test old ones. Megx.net is a technology that helps us answer the questions we have and to keep looking for new ones.

CHAPTER 5

CONCLUSION

Marine ecological genomics will continue to benefit greatly from improvements to sequencing technologies and application of multiomics approaches. Integrating the resulting data and complementing it with ecological metadata will remain the main challenge for the next decade. Megx.net delivers a platform for robust environmental data integration which can be used to investigate the interaction of microbial communities with their environment.

Appendix

Additional Scientific Publications

A list of scientific publications that resulted form the work in this thesis but were not dicussed in detail.

1. **CDinFusion - Submission-ready, on-line Integration of Sequence and Contextual Data**
   Authors: Wolfgang Hankeln, Norma Wendel, Jan Gerken, Jost Waldmann, Pier Luigi Buttigieg, Ivaylo Kostadinov, Renzo Kottmann, Pelin Yilmaz, Frank Oliver Glöckner
   Submitted to: PLoS One
   Contribution: Web Services
   Description: Describes a web-based tool for adding GSC compliant metadata to sequence FASTA files, which can then be readily submitted to the public databases.

2. **Metagenome and mRNA expression analyses of anaerobic methanotrophic archaea of the ANME-1 group**
   Authors: Anke Meyerdierks, Michael Kube, Ivaylo Kostadinov, Hanno Teeling, Frank Oliver Glöckner, Richard Reinhardt, Rudolf Amann
   Published in: Environmental Microbiology, (doi:10.1111/j.1462-2920.2009.02083.x)
   Contribution: estimation of genome coverage using single-copy genes
   Description: Construction and genomic analysis of a compostie genome of methanotrophic archaea fromthe ANME-1 clade.

Acknowledgements

I would like to express my sincerest gratitude to Prof. Dr. Frank Oliver Glöckner. He opened the door to Bioinformatics for me and helped me to take the first step, and the second, and the third... I owe him a lot for the supervision and absolute support during my complete higher education. Naturally, I am most grateful for the opportunity to conduct this PhD work in his group at the Max Planck Institute for Marine Microbiology. He always went out of his ways to make me feel comfortable.

Next, I would like to thank my thesis committee members, Prof. Dr. Frank Oliver Glöckner, PD Dr. Bernhard Fuchs and Prof. Dr. Matthias Ullrich. They have been very supportive of my work and have contributed valuable insights from their own experience and fields. It was very important for me to receive their positive opinion of my work, which I did.

The Microbial Genomics Group at the Max Planck Institute for Marine Microbiology has been the best working environment I could imagine. All members (current and past) together and in individual are responsible for this. I would like to especially thank Dr. Renzo Kottmann, from whom I have learned the most in the past years and Dr. Melissa Duhaime, who brought viruses in our boring microbial life ;) And, of course, Dr. Michael Richter for *Schefflera arboricola*. Our office is the best! I want to thank the rest of the megx.net team: Pier Luigi Buttigieg, Pelin Yilmaz and Wolfgang Hankeln, for the and their devotion to the project. I want to thank Jost Waldmann for various (heated) discussions, for winning the MPI Foosball cup and for his friendship ;)

The Max Planck Research School of Marine Microbiology Program at the Max Planck Institute for Marine Microbiology, especially Christiane Glöckner. I believe she is known as the 'substitute mom', for she has

been something like a mother to us all. The faculty members and students, which made the first one and a half years an experience worth telling about. A very special thanks to my MarMic class of 2009 for that very special bond that emerged between us from nowhere (probably Sylt) and will be never severed. Another special thanks to MarMic class 2010, for socially adopting me in a way people thought I am part of the class. You are great! Of course, my "boyz frm dahud" get a special wink for everything...
All in all, the Max Planck Institute for Marine Microbiology and the MarMic Program have given me the best 4 years of my life concerning both academic and private life. I was able to make amazing friends with my colleagues. That is exceptional! Thank you!
A very special thanks goes to my family: my brother Momchil, my mother Galina, and her parents Ivan and Svetla, for their unconditional love and support over the last 27 years. Thank you so much! I love you from all my heart!
One person had to put up with all my ups and downs during my PhD. She did so magnificently and even managed to stay in love with me.
Liebe Jessi, dir auch, vielen Herzlichen Dank!

Bibliography

[Amann et al., 1990] Amann, R. I., Krumholz, L., and Stahl, D. A. (1990). Fluorescent-oligonucleotide probing of whole cells for determinative, phylogenetic, and environmental studies in microbiology. *Journal of Bacteriology*, 172(2):762–770.

[Amann et al., 1995] Amann, R. I., Ludwig, W., and Schleifer, K. H. (1995). Phylogenetic identification and in situ detection of individual microbial cells without cultivation. *Microbiological Reviews*, 59(1):143–169.

[Arita, 2009] Arita, M. (2009). A pitfall of wiki solution for biological databases. *Briefings in Bioinformatics*, 10(3):295–296.

[Béjà et al., 2000] Béjà, O., Aravind, L., Koonin, E. V., Suzuki, M. T., Hadd, A., Nguyen, L. P., Jovanovich, S. B., Gates, C. M., Feldman, R. A., Spudich, J. L., Spudich, E. N., and DeLong, E. F. (2000). Bacterial rhodopsin: evidence for a new type of phototrophy in the sea. *Science*, 289(5486):1902–1906.

[Berg, 1996] Berg, R. (1996). The indigenous gastrointestinal microflora. *Trends in Microbiology*, 4(11):430–435.

[Beszteri et al., 2010] Beszteri, B., Temperton, B., Frickenhaus, S., and Giovannoni, S. J. (2010). Average genome size: a potential source of bias in comparative metagenomics. *The ISME Journal*, 4(8):1075–7.

[Binga et al., 2008] Binga, E. K., Lasken, R. S., and Neufeld, J. D. (2008). Something from (almost) nothing: the impact of multiple displacement amplification on microbial ecology. *The ISME Journal*, 2(3):233–241.

[Boyer et al., 2006] Boyer, T. P., Antonov, J., Garcia, H., Johnson, D., Locarnini, R., Mishonov, A., Pitcher, M., Baranova, O., and Smolyar, I. (2006). World Ocean Database 2005. Technical report, S. Levitus, Ed., NOAA Atlas NESDIS 60, U.S. Government Printing Office, Washington, D.C., 190 pp., DVDs.

[Cases et al., 2003] Cases, I., de Lorenzo, V., and Ouzounis, C. A. (2003). Transcription regulation and environmental adaptation in bacteria. *Trends in Microbiology*, 11(6):248–253.

[Croucher, 2009] Croucher, N. J. (2009). From small reads do mighty genomes grow. *Nature Reviews Microbiology*, 7(9):621.

[Dahm, 2007] Dahm, R. (2007). Discovering DNA: Friedrich Miescher and the early years of nucleic acid research. *Human Genetics*, 122(6):565–581.

[DeLong et al., 1989] DeLong, E. F., Wickham, G. S., and Pace, N. R. (1989). Phylogenetic stains: ribosomal RNA-based probes for the identification of single cells. *Science*, 243(4896):1360–1363.

[Eddy, 2009] Eddy, S. R. (2009). A new generation of homology search tools based on probabilistic inference. *Genome Informatics. International Conference on Genome Informatics*, 23(1):205–211.

[Editorial, 2009] Editorial (2009). Metagenomics versus Moore's law. *Nature Methods*, 6(9):623–623.

[Field, 1998] Field, C. B. (1998). Primary Production of the Biosphere: Integrating Terrestrial and Oceanic Components. *Science*, 281(5374):237–240.

[Field, 2008] Field, D. (2008). Working together to put molecules on the map. *Nature*, 453(7198):978.

[Field et al., 2008] Field, D., Garrity, G., Gray, T., Morrison, N., Selengut, J., Sterk, P., Tatusova, T., Thomson, N., Allen, M. J., Angiuoli, S. V., 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, F. O., 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, S. E., Li, K., Lister, A. L., Lord, P., Maltsev, N., Markowitz, V., Martiny, J., Methe, B., Mizrachi, I., Moxon, R., Nelson, K., Parkhill, J., Proctor, L., White, O., Sansone, S.-A., 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., and Wipat, A. (2008). The minimum information about a genome sequence (MIGS) specification. *Nature Biotechnology*, 26(5):541–7.

[Fiore et al., 2010] Fiore, C. L., Jarett, J. K., Olson, N. D., and Lesser, M. P. (2010). Nitrogen fixation and nitrogen transformations in marine symbioses. *Trends in Microbiology*, 18(10):455–463.

[Fleischmann et al., 1995] Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., and Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, 269(5223):496–512.

[Francis et al., 2007] Francis, C. A., Beman, J. M., and Kuypers, M. M. M. (2007). New processes and players in the nitrogen cycle: the microbial ecology of anaerobic and archaeal ammonia oxidation. *The ISME Journal*, 1(1):19–27.

[Frias-Lopez et al., 2008] Frias-Lopez, J., Shi, Y., Tyson, G. W., Coleman, M. L., Schuster, S. C., Chisholm, S. W., and Delong, E. F. (2008). Microbial community gene expression in ocean surface waters. *PNAS*, 105(10):3805–10.

[Gianoulis et al., 2009] Gianoulis, T. A., Raes, J., Patel, P. V., Bjornson, R., Korbel, J. O., Letunic, I., Yamada, T., Paccanaro, A., Jensen, L. J., Snyder, M., Bork, P., and Gerstein, M. B. (2009). Quantifying environmental adaptation of metabolic pathways in metagenomics. *PNAS*, 106(5):1374–9.

[Gilbert et al., 2010] Gilbert, J. A., Field, D., Swift, P., Thomas, S., Cummings, D., Temperton, B., Weynberg, K., Huse, S., Hughes, M.,

Joint, I., Somerfield, P. J., and 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(11).

[Goble and Stevens, 2008] Goble, C. and Stevens, R. (2008). State of the nation in data integration for bioinformatics. *Journal of Biomedical Informatics*, 41(5):687–693.

[Goble et al., 2008] Goble, C., Stevens, R., Hull, D., Wolstencroft, K., and Lopez, R. (2008). Data curation + process curation=data integration + science. *Briefings in Bioinformatics*, 9(6):506–517.

[Goll et al., 2010] Goll, J., Rusch, D., Tanenbaum, D. M., Thiagarajan, M., Li, K., Methé, B. A., and Yooseph, S. (2010). METAREP: JCVI Metagenomics Reports - an open source tool for high-performance comparative metagenomics. *Bioinformatics*, 26(20):2631–2632.

[Grossart et al., 2005] Grossart, H.-P., Levold, F., Allgaier, M., Simon, M., and Brinkhoff, T. (2005). Marine diatom species harbour distinct bacterial communities. *Environmental Microbiology*, 7(6):860–873.

[Handelsman, 2004] Handelsman, J. (2004). Metagenomics: Application of Genomics to Uncultured Microorganisms. *Microbiology and Molecular Biology Reviews*, 68(4):669–685.

[Handelsman et al., 1998] Handelsman, J., Rondon, M. R., Brady, S. F., Clardy, J., and Goodman, R. M. (1998). Molecular biological access to the chemistry of unknown soil microbes: a new frontier for natural products. *Chemistry & Biology*, 5(10):R245–249.

[Hankeln et al., 2010] Hankeln, W., Buttigieg, P. L., Fink, D., Kottmann, R., Yilmaz, P., and Glöckner, F. O. (2010). MetaBar - a tool for consistent contextual data acquisition and standards compliant submission. *BMC bioinformatics*, 11(1):358–366.

[Hankeln et al.,] Hankeln, W., Buttigieg, P. L., Gerken, J., Kostadinov, I., Kottmann, R., Waldmann, J., Wendel, N., Yilmaz, P., and Glöckner, F. O. CDinFusion - Submission-ready, on-line Integration of Sequence and Contextual Data. under rev.

[Hingamp et al., 2008] Hingamp, P., Brochier, C., Talla, E., Gautheret, D., Thieffry, D., and Herrmann, C. (2008). Metagenome Annotation Using a Distributed Grid of Undergraduate Students. *PLoS Biology*, 6(11):e296.

[Hirschman et al., 2008] Hirschman, L., Clark, C., Cohen, K. B., Mardis, S., Luciano, J., Kottmann, R., Cole, J., Markowitz, V., Kyrpides, N., Morrison, N., Schriml, L. M., and Field, D. (2008). HabitatLite: a GSC case study based on free text terms for environmental metadata. *Omics : a journal of integrative biology*, 12(2):129–36.

[Huber et al., 2000] Huber, R., Huber, H., and Stetter, K. O. (2000). Towards the ecology of hyperthermophiles: biotopes, new isolation strategies and novel metabolic properties. *FEMS Microbiology Reviews*, 24(5):615–623.

[Hugenholtz and Tyson, 2008] Hugenholtz, P. and Tyson, G. W. (2008). Microbiology: Metagenomics. *Nature*, 455(7212):481–483.

[JOU et al., 1972] JOU, W. M., HAEGEMAN, G., YSEBAERT, M., and FIERS, W. (1972). Nucleotide Sequence of the Gene Coding for the Bacteriophage MS2 Coat Protein. *Nature*, 237(5350):82–88.

[Kalyuzhnaya et al., 2008] Kalyuzhnaya, M. G., Lapidus, A., Ivanova, N., Copeland, A. C., McHardy, A. C., Szeto, E., Salamov, A., Grigoriev, I. V., Suciu, D., Levine, S. R., Markowitz, V. M., Rigoutsos, I., Tringe, S. G., Bruce, D. C., Richardson, P. M., Lidstrom, M. E., and Chistoserdova, L. (2008). High-resolution metagenomics targets specific functional types in complex microbial communities. *Nature Biotechnology*, 26(9):1029–34.

[Karaoz et al., 2004] Karaoz, U., Murali, T. M., Letovsky, S., Zheng, Y., Ding, C., Cantor, C. R., and Kasif, S. (2004). Whole-genome annotation by using evidence integration in functional-linkage networks. *PNAS*, 101(9):2888–93.

[Kivistö and Karp, 2010] Kivistö, A. T. and Karp, M. T. (2010). Halophilic anaerobic fermentative bacteria. *Journal of Biotechnology*, 152(4):114–124.

[Klappenbach et al., 2000] Klappenbach, J. A., Dunbar, J. M., and Schmidt, T. M. (2000). rRNA Operon Copy Number Reflects Ecological Strategies of Bacteria. *Applied and Environmental Microbiology*, 66(4):1328–1333.

[Kostadinov et al.,] Kostadinov, I., Kottmann, R., Ramette, A., Waldmann, J., Buttigieg, P. L., and Glöckner, F. O. Quantifying the Effect of Environment Stability on the Transcription Factor Repertoire of Marine Microbes. under rev.

[Kottmann et al., 2008] Kottmann, R., Gray, T., Murphy, S., Kagan, L., Kravitz, S., Lombardot, T., Field, D., and Glöckner, F. O. (2008). A standard MIGS/MIMS compliant XML Schema: toward the development of the Genomic Contextual Data Markup Language (GCDML). *Omics: A Journal of Integrative Biology*, 12(2):115–121.

[Kottmann et al., 2010] Kottmann, R., Kostadinov, I., Duhaime, M. B., Buttigieg, P. L., Yilmaz, P., Hankeln, W., Waldmann, J., and Glöckner, F. O. (2010). Megx.net: integrated database resource for marine ecological genomics. *NAR*, 38(Database issue):D391–395.

[Lauro et al., 2009] Lauro, F. M., McDougald, D., Thomas, T., Williams, T. J., Egan, S., Rice, S., DeMaere, M. Z., Ting, L., Ertan, H., Johnson, J., Ferriera, S., Lapidus, A., Anderson, I., Kyrpides, N., Munk, A. C., Detter, C., Han, C. S., Brown, M. V., Robb, F. T., Kjelleberg, S., and Cavicchioli, R. (2009). The genomic basis of trophic strategy in marine bacteria. *PNAS*, 106(37):15527–15533.

[Liolios et al., 2010] Liolios, K., Chen, I.-M. A., Mavromatis, K., Tavernarakis, N., Hugenholtz, P., Markowitz, V. M., and Kyrpides, N. C. (2010). The Genomes On Line Database (GOLD) in 2009: status of genomic and metagenomic projects and their associated metadata. *NAR*, 38(Database issue):D346–354.

[Liu et al., 2007] Liu, Z., Lozupone, C., Hamady, M., Bushman, F. D., and Knight, R. (2007). Short pyrosequencing reads suffice for accurate microbial community analysis. *NAR*, 35(18):e120.

[Lombardot et al., 2006] Lombardot, T., Kottmann, R., Pfeffer, H., Richter, M., Teeling, H., Quast, C., and Glöckner, F. O. (2006).

Megx.net–database resources for marine ecological genomics. *NAR*, 34(Database issue):D390–393.

[Madigan et al., 2003] Madigan, M., Martinko, J., and Parker, J., editors (2003). *Brock Biology of Microorganisms, 10th edition*.

[Manavski and Valle, 2008] Manavski, S. A. and Valle, G. (2008). CUDA compatible GPU cards as efficient hardware accelerators for Smith-Waterman sequence alignment. *BMC bioinformatics*, 9(Suppl 2):S10–19.

[Markowitz et al., 2008] Markowitz, V. M., Ivanova, N. N., Szeto, E., Palaniappan, K., Chu, K., Dalevi, D., Chen, I.-M. A., Grechkin, Y., Dubchak, I., Anderson, I., Lykidis, A., Mavromatis, K., Hugenholtz, P., and Kyrpides, N. C. (2008). IMG/M: a data management and analysis system for metagenomes. *NAR*, 36(Database issue):D534–538.

[Martin et al., 2008] Martin, C., Diaz, N. N., Ontrup, J., and Nattkemper, T. W. (2008). Hyperbolic SOM-based clustering of DNA fragment features for taxonomic visualization and classification. *Bioinformatics*, 24(14):1568–1574.

[McCarren et al., 2010] McCarren, J., Becker, J. W., Repeta, D. J., Shi, Y., Young, C. R., Malmstrom, R. R., Chisholm, S. W., and DeLong, E. F. (2010). Inaugural Article: Microbial community transcriptomes reveal microbes and metabolic pathways associated with dissolved organic matter turnover in the sea. *PNAS*, 107(38):16420–16427.

[Meinicke, 2009] Meinicke, P. (2009). UFO: a web server for ultra-fast functional profiling of whole genome protein sequences. *BMC Genomics*, 10:409–418.

[Metzker, 2009] Metzker, M. L. (2009). Sequencing technologies — the next generation. *Nature Reviews Genetics*, 11(1):31–46.

[Meyer et al., 2008] Meyer, F., Paarmann, D., D'Souza, M., Olson, R., Glass, E., Kubal, M., Paczian, T., Rodriguez, A., Stevens, R., Wilke, A., Wilkening, J., and Edwards, R. (2008). The metagenomics RAST server - a public resource for the automatic phylogenetic and functional analysis of metagenomes. *BMC Bioinformatics*, 9(1):386.

[Meyerdierks et al., 2010] Meyerdierks, A., Kube, M., Kostadinov, I., Teeling, H., Glöckner, F. O., Reinhardt, R., and Amann, R. (2010). Metagenome and mRNA expression analyses of anaerobic methanotrophic archaea of the ANME-1 group. *Environmental Microbiology*, 12(2):422–39.

[Morris et al., 2010] Morris, R. M., Nunn, B. L., Frazar, C., Goodlett, D. R., Ting, Y. S., and Rocap, G. (2010). Comparative metaproteomics reveals ocean-scale shifts in microbial nutrient utilization and energy transduction. *The ISME Journal*, 4(5):673–85.

[Pace et al., 1985] Pace, N., Stahl, D., Lane, D., and Olsen, G. (1985). Analyzing natural microbial populations by rRNA sequences. *American Society for Microbiology News*, 51(1):4–12.

[Pace, 1997] Pace, N. R. (1997). A molecular view of microbial diversity and the biosphere. *Science*, 276(5313):734–740.

[Palenik et al., 2006] Palenik, B., Ren, Q., Dupont, C. L., Myers, G. S., Heidelberg, J. F., Badger, J. H., Madupu, R., Nelson, W. C., Brinkac, L. M., Dodson, R. J., Durkin, A. S., Daugherty, S. C., Sullivan, S. A., Khouri, H., Mohamoud, Y., Halpin, R., and Paulsen, I. T. (2006). Genome sequence of Synechococcus CC9311: Insights into adaptation to a coastal environment. *PNAS*, 103(36):13555–9.

[Parks and Beiko, 2010] Parks, D. H. and Beiko, R. G. (2010). Identifying biologically relevant differences between metagenomic communities. *Bioinformatics*, 26(6):715–721.

[Pruesse et al., 2007] Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., and Glöckner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. *NAR*, 35(21):7188–7196.

[Rappé and Giovannoni, 2003] Rappé, M. S. and Giovannoni, S. J. (2003). The uncultured microbial majority. *Annual Review of Microbiology*, 57:369–394.

[Raven J. A. and Falkowski, 1999] Raven J. A. and Falkowski, P. G. (1999). Oceanic sinks for atmospheric CO 2. *Plant, Cell and Environment*, 22(6):741–755.

[Riesenfeld et al., 2004] Riesenfeld, C. S., Schloss, P. D., and Handelsman, J. (2004). METAGENOMICS: Genomic Analysis of Microbial Communities. *Annual Review of Genetics*, 38(1):525–552.

[Rusch et al., 2007] Rusch, D. B., Halpern, A. L., Sutton, G., Heidelberg, K. B., Williamson, S., Yooseph, S., Wu, D., Eisen, J. A., Hoffman, J. M., Remington, K., Beeson, K., Tran, B., Smith, H., Baden-Tillson, H., Stewart, C., Thorpe, J., Freeman, J., Andrews-Pfannkoch, C., Venter, J. E., Li, K., Kravitz, S., Heidelberg, J. F., Utterback, T., Rogers, Y.-H., Falcón, L. I., Souza, V., Bonilla-Rosso, G., Eguiarte, L. E., Karl, D. M., Sathyendranath, S., Platt, T., Bermingham, E., Gallardo, V., Tamayo-Castillo, G., Ferrari, M. R., Strausberg, R. L., Nealson, K., Friedman, R., Frazier, M., and Venter, J. C. (2007). The Sorcerer II Global Ocean Sampling expedition: northwest Atlantic through eastern tropical Pacific. *PLoS Biology*, 5(3):e77.

[Savage, 1977] Savage, D. C. (1977). Microbial ecology of the gastrointestinal tract. *Annual Review of Microbiology*, 31(1):107–133.

[Schadt et al., 2010] Schadt, E. E., Linderman, M. D., Sorenson, J., Lee, L., and Nolan, G. P. (2010). Computational solutions to large-scale data management and analysis. *Nature Reviews Genetics*, 11(9):647–657.

[Schattenhofer et al., 2009] Schattenhofer, M., Fuchs, B. M., Amann, R., Zubkov, M. V., Tarran, G. A., and Pernthaler, J. (2009). Latitudinal distribution of prokaryotic picoplankton populations in the Atlantic Ocean. *Environmental Microbiology*, 11(8):2078–2093.

[Schattenhofer et al., 2011] Schattenhofer, M., Wulf, J., Kostadinov, I., Glöckner, F. O., Zubkov, M. V., and Fuchs, B. M. (2011). Phylogenetic Characterisation of Picoplanktonic Populations with High and Low Nucleic Acid Content in the North Atlantic Ocean. *Systematic and Applied Microbiology*, in press.

[Schloss and Handelsman, 2008] Schloss, P. and Handelsman, J. (2008). A statistical toolbox for metagenomics: assessing functional diversity in microbial communities. *BMC Bioinformatics*, 9(1):34–49.

[Schulz et al., 1999] Schulz, H. N., Brinkhoff, T., Ferdelman, T. G., Mariné, M. H., Teske, A., and Jø rgensen, B. B. (1999). Dense Populations of a Giant Sulfur Bacterium in Namibian Shelf Sediments. *Science*, 284(5413):493 –495.

[Schulz and Jorgensen, 2001] Schulz, H. N. and Jorgensen, B. B. (2001). BIG BACTERIA. *Annual Review of Microbiology*, 55(1):105–137.

[Seshadri et al., 2007] Seshadri, R., Kravitz, S. A., Smarr, L., Gilna, P., and Frazier, M. (2007). CAMERA: A Community Resource for Metagenomics. *PLoS Biology*, 5(3):e75.

[Shendure and Ji, 2008] Shendure, J. and Ji, H. (2008). Next-generation DNA sequencing. *Nature Biotechnology*, 26(10):1135–1145.

[Simonato et al., 2010] Simonato, F., Gómez-Pereira, P. R., Fuchs, B. M., and Amann, R. (2010). Bacterioplankton diversity and community composition in the Southern Lagoon of Venice. *Systematic and applied microbiology*, 33(3):128–38.

[Smetacek and Naqvi, 2008] Smetacek, V. and Naqvi, S. W. A. (2008). The next generation of iron fertilization experiments in the Southern Ocean. *Philosophical transactions. Series A, Mathematical, physical, and engineering sciences*, 366(1882):3947–3967.

[Sowell et al., 2011] Sowell, S. M., Abraham, P. E., Shah, M., Verberkmoes, N. C., Smith, D. P., Barofsky, D. F., and Giovannoni, S. J. (2011). Environmental proteomics of microbial plankton in a highly productive coastal upwelling system. *The ISME journal*, 5:856–865.

[Stein et al., 1996] Stein, J. L., Marsh, T. L., Wu, K. Y., Shizuya, H., and DeLong, E. F. (1996). Characterization of uncultivated prokaryotes: isolation and analysis of a 40-kilobase-pair genome fragment from a planktonic marine archaeon. *Journal of Bacteriology*, 178(3):591–599.

[Stein, 2010] Stein, L. (2010). The case for cloud computing in genome informatics. *Genome Biology*, 11(5):207–214.

[Stratton et al., 2009] Stratton, M. R., Campbell, P. J., and Futreal, P. A. (2009). The cancer genome. *Nature*, 458(7239):719–724.

[Sun et al., 2009] Sun, Y., Cai, Y., Liu, L., Yu, F., Farrell, M. L., McKendree, W., and Farmerie, W. (2009). ESPRIT: estimating species richness using large collections of 16S rRNA pyrosequences. *Nucleic acids research*, 37(10):e76.

[Suttle, 2005] Suttle, C. A. (2005). Viruses in the sea. *Nature*, 437(7057):356–61.

[Suttle, 2007] Suttle, C. A. (2007). Marine viruses–major players in the global ecosystem. *Nature Reviews Microbiology*, 5(10):801–12.

[Tripp et al., 2010] Tripp, H. J., Bench, S. R., Turk, K. A., Foster, R. A., Desany, B. A., Niazi, F., Affourtit, J. P., and Zehr, J. P. (2010). Metabolic streamlining in an open-ocean nitrogen-fixing cyanobacterium. *Nature*, 464(7285):90–94.

[Tyson et al., 2004] Tyson, G. W., Chapman, J., Hugenholtz, P., Allen, E. E., Ram, R. J., Richardson, P. M., Solovyev, V. V., Rubin, E. M., Rokhsar, D. S., and Banfield, J. F. (2004). Community structure and metabolism through reconstruction of microbial genomes from the environment. *Nature*, 428(6978):37–43.

[van Straalen and Roelofs, 2006] van Straalen, N. M. and Roelofs, D. (2006). *An Introduction to Ecological Genomics*. Oxford University Press.

[Venter et al., 2004] Venter, J. C., Remington, K., Heidelberg, J. F., Halpern, A. L., Rusch, D., Eisen, J. A., Wu, D., Paulsen, I., Nelson, K. E., Nelson, W., Fouts, D. E., Levy, S., Knap, A. H., Lomas, M. W., Nealson, K., White, O., Peterson, J., Hoffman, J., Parsons, R., Baden-Tillson, H., Pfannkoch, C., Rogers, Y.-H., and Smith, H. O. (2004). Environmental genome shotgun sequencing of the Sargasso Sea. *Science*, 304(5667):66–74.

[Warnecke and Hugenholtz, 2007] Warnecke, F. and Hugenholtz, P. (2007). Building on basic metagenomics with complementary technologies. *Genome Biology*, 8(12):231–236.

[Weber et al., 2011] Weber, M., Teeling, H., Huang, S., Waldmann, J., Kassabgy, M., Fuchs, B. M., Klindworth, A., Klockow, C., Wichels, A., Gerdts, G., Amann, R., and Glöckner, F. O. (2011). Practical application of self-organizing maps to interrelate biodiversity and functional data in NGS-based metagenomics. *The ISME journal*, 5:918–928.

[Whitman et al., 1998] Whitman, W. B., Coleman, D. C., and Wiebe, W. J. (1998). Prokaryotes: the unseen majority. *PNAS*, 95(12):6578–6583.

[Wilkening et al., 2009] Wilkening, J., Wilke, A., Desai, N., and Meyer, F. (2009). Using clouds for metagenomics: A case study. In *2009 IEEE International Conference on Cluster Computing and Workshops*, pages 1–6. IEEE.

[Woodward, 2007] Woodward, F. I. (2007). Global primary production. *Current biology : CB*, 17(8):R269–273.

[Woyke et al., 2009] Woyke, T., Xie, G., Copeland, A., González, J. M., Han, C., Kiss, H., Saw, J. H., Senin, P., Yang, C., Chatterji, S., Cheng, J.-F., Eisen, J. A., Sieracki, M. E., and Stepanauskas, R. (2009). Assembling the Marine Metagenome, One Cell at a Time. *PloS One*, 4(4):e5299.

[Wu et al., 2009] Wu, D., Hugenholtz, P., Mavromatis, K., Pukall, R., Dalin, E., Ivanova, N. N., Kunin, V., Goodwin, L., Wu, M., Tindall, B. J., Hooper, S. D., Pati, A., Lykidis, A., Spring, S., Anderson, I. J., D'haeseleer, P., Zemla, A., Singer, M., Lapidus, A., Nolan, M., Copeland, A., Han, C., Chen, F., Cheng, J.-F., Lucas, S., Kerfeld, C., Lang, E., Gronow, S., Chain, P., Bruce, D., Rubin, E. M., Kyrpides, N. C., Klenk, H.-P., and Eisen, J. A. (2009). A phylogeny-driven genomic encyclopaedia of Bacteria and Archaea. *Nature*, 462(7276):1056–1060.

[Yilmaz et al., 2011a] Yilmaz, P., Gilbert, J. A., Knight, R., Amaral-Zettler, L., Karsch-Mizrachi, I., Cochrane, G., Nakamura, Y., Sansone, S.-A., Glöckner, F. O., and Field, D. (2011a). The genomic standards consortium: bringing standards to life for microbial ecology. *The ISME Journal*, online.

[Yilmaz et al., 2011b] Yilmaz, P., Kottman, R., Field, D., Knight, R., Cole, J., Amaral-Zettler, L., and Al., E. (2011b). "Minimum Information about a MARKer gene Sequence" (MIMARKS) specification. *Nature Biotechnology*, in press.

[Yooseph et al., 2010] Yooseph, S., Nealson, K. H., Rusch, D. B., McCrow, J. P., Dupont, C. L., Kim, M., Johnson, J., Montgomery, R., Ferriera, S., Beeson, K., Williamson, S. J., Tovchigrechko, A., Allen, A. E., Zeigler, L. A., Sutton, G., Eisenstadt, E., Rogers, Y.-H., Friedman, R., Frazier, M., and Venter, J. C. (2010). Genomic and functional adaptation in surface ocean planktonic prokaryotes. *Nature*, 468(7320):60–66.

[Zhang et al., 2010] Zhang, W., Li, F., and Nie, L. (2010). Integrating multiple 'omics' analysis for microbial biology: application and methodologies. *Microbiology*, 156:287–301.

[Zhang et al., 2009] Zhang, Z., Cheung, K.-H., and Townsend, J. P. (2009). Bringing Web 2.0 to bioinformatics. *Briefings in Bioinformatics*, 10(1):1–10.